UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Speech as Biomarker for Multidisease Screening

Maria Catarina Tavares Botelho

Supervisor: Doctor Isabel Maria Martins Trancoso
Co-supervisors: Doctor Alberto Abad Gareta
Doctor Tanja Schultz

Thesis approved in public session to obtain the PhD Degree in Electrical and Computer Engineering

Jury final classification: Pass with Distinction and Honour

2024

TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Speech as Biomarker for Multidisease Screening

Maria Catarina Tavares Botelho

Supervisor: Doctor Isabel Maria Martins Trancoso
Co-supervisors: Doctor Alberto Abad Gareta
Doctor Tanja Schultz

Thesis approved in public session to obtain the PhD Degree in Electrical and Computer Engineering

Jury final classification: Pass with Distinction and Honour

Jury

Chairperson: Doctor Rui Jorge Morais Tomaz Valadas, Instituto Superior Técnico, Universidade de Lisboa

Members of the Committee:
Doctor Helen Meng, Faculty of Engineering, The Chinese University of Hong Kong, Hong Kong
Doctor Tanja Schultz, Faculty 3 - Mathematics and Computer Science, University of Bremen, Alemanha
Doctor Maria Luísa Torres Ribeiro Marques da Silva Coheur, Instituto Superior Técnico, Universidade de Lisboa
Doctor André Filipe Torres Martins, Instituto Superior Técnico, Universidade de Lisboa
Doctor Nicholas Cummins, Institute of Psychiatry, Psychology & Neuroscience, School of Mental Health & Psychological Sciences, Department of Biostatistics and Health Informatics, King's College London, Reino Unido
Doctor Rui Miguel Carrasqueiro Henriques, Instituto Superior Técnico, Universidade de Lisboa

Funding Institution
FCT: Fundação para a Ciência e a Tecnologia
Portuguese Recovery and Resilience Plan and NextGenerationEU European Union Funds

2024

Para os meus tios Artur e Carlos, com esperança.

Acknowledgments

First and foremost, my deepest gratitude goes to Professor Isabel and Professor Alberto. They welcomed me in 2018 to conduct my MSc Thesis, and they have continued to welcome and guide me every day since. I thank them for the technical support and scientific discussions, but also for their kindness, their humanity, and their friendship. I am especially grateful for their willingness to always make time for me, often late into the night. I also wish to extend my heartfelt thanks to Professor Tanja, who accepted to be my remote supervisor, providing guidance, insightful discussions, and numerous opportunities. I thank her for welcoming me to Bremen and treating me as part of the CSL family. I am grateful to my advisors for their encouragement to reach my full potential, for the freedom they gave me to explore my ideas, and for always asking the right questions. There are not enough words to express my gratitude towards them, except to say I could never have chosen better advisors.

I extend my thanks to the members of my examination committee who have taken the time to read and discuss this thesis. Furthermore, I especially thank the committee members who participated in the thesis proposal for their invaluable feedback and for ensuring that this thesis followed the right track.

I would also like to acknowledge the support from Fundação para a Ciência e Tecnologia and the PRR funds that made this work possible[1].

My gratitude goes also to everyone at the Human Language Technologies Lab, for creating an ideal working atmosphere that fosters scientific collaboration and discussion, as well as dinners, cake, and padel matches. Particularly, I would like to give a special acknowledgment to my academic siblings – Francisco, Joana, Mariana, Thomas, Carlos, and John – for always being ready to help, for the breakfasts, lunches, and coffee breaks, for the travels, for the conversations, for making every day as a PhD student better. I thank Anna for her collaboration in the past months, as well as for always having an optimistic perspective on everything. I also extend my thanks to all other PhD students, but also to Rubén, Lena, Luísa, and Fernando. I also thank David for maintaining the infrastructure and always quickly responding to all our inquiries.

Additionally, I am grateful to the folks at CSL, especially Ay, for being a great colleague and teaching

[1]This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia, with references DOI: 10.54499/UIDB/50021/2020 and SFRH/BD/149126/2019, and by the Portuguese Recovery and Resilience Plan and Next Generation EU European Funds, through project C644865762- 00000008 (Accelerat.AI).

me so much about life, courage, and strength, always with a smile. I also thank Rathi and Kevin for making my stay in Bremen so enjoyable.

I would like to thank Vaiva, my supervisor during my internship at Google AI, for her technical support and code reviews, and for being an inspiration on balancing life and work.

I also thank all the colleagues and friends I met at ISCA-SAC and the ISCA Mentoring Committee. It was a pleasure to work with you and learn the meaning of community in research. I believe the work we have conducted together has significantly improved the experience of many PhD students in our community, and this achievement is just as important as the findings described in this thesis.

I would like to thank my friends who have accompanied me since I entered Técnico in 2013. I would also like to thank my friends from ballet class, and my life-long ballet teacher. I would like to thank my friends from my high school and even kindergarten who continue my friends today. They all keep believing in me, supporting me, and making sure I keep mentally sane.

I would like to thank my entire family, I am very grateful for the support and the confidence they have placed in me, but also for the constant laughter we share. Special thanks to my parents, who instilled in me a passion for science and a curiosity to question everything long before I began my degree at Técnico. I thank my mother for being a true inspiration, for being "*uma das inderrotáveis*"[2]. When I grow up, I aspire to be like her. I thank my father for always being available to discuss everything, from major life decisions to quantum physics and neuroscience concepts. I thank him for having the time to read all my work (and enjoying to do so), and for being able to teach me everything. To my brother, who knows me like no one else, who only sees the best in everyone and everything, and who knows exactly how to make me feel good at all times. I also thank my cousins for transforming my mother's house into a student dorm and contributing to making it the happiest place on earth.

Finally, my profound thanks to Manel for always listening, whether it be about the tiniest indecision or the largest problems, for accepting to hug me when I am not ready to talk, and for all his patience. Above all, I thank him for supporting all my craziest ideas, and adding colour to them.

This PhD is mine, but it is also theirs.

[2] O elogio das inderrotáveis, Miguel Esteves Cardoso, Crónica do Público, 11 de Maio de 2023

Abstract

Overburdened health systems worldwide face challenges exacerbated by an aging population. Speech, a rich and ubiquitous biomarker, offers the potential for a widespread low-cost detection of neurodegenerative, psychiatric, and respiratory diseases. This potential stems from the involvement of the respiratory, nervous, and muscular systems in speech production, which encodes information on dysfunctions in any of these systems. At all levels of speech production, biosignals can be captured and studied to obtain paralinguistic information.

This thesis begins by exploring one such biosignal–electromyography signals (EMG) produced during speech articulation–establishing the foundation for Silent Computational Paralinguistics. During COVID-19, remote biomarkers gained importance. We explored facial images and visual speech combined with acoustic speech to detect obstructive sleep apnea, achieving promising results with knowledge-based and transfer learning methods in a pilot study of 40 subjects.

Despite high performance in the automatic detection of speech-affecting diseases, questions remain about what these models are actually learning and the basis for their predictions, which can significantly impact patients' lives. Ensuring the reliability and generalizability of the results is crucial. We advocate for a robust and interpretable health monitoring model, suitable for the simultaneous detection of several diseases, as speech-affecting disorders often have overlapping effects on the speech signal. We propose a framework for defining normative speech through reference intervals of clinically significant features. We leverage deviations from this model to perform the detection of Alzheimer's and Parkinson's diseases, using different classifiers, namely Neural Additive Models for enhanced interpretability.

Furthermore, in a quest to bridge black-box models and interpretability, we explore large language models to annotate high-level, low-dimensional interpretable speech characteristics, termed macro-descriptors (e.g., text coherence, lexical diversity). Using only four macro-descriptors, we outperform Alzheimer's detection with conventional language-based features.

This thesis contributes to a deeper understanding of the multifaceted potential of speech as a biomarker for holistic health.

Keywords

Speech; Speech affecting diseases; Multimorbidty; Multimodality; Interpretability

Resumo

Atualmente, os sistemas de saúde estão sobrecarregados em todo o mundo e enfrentam enormes desafios agravados pelo envelhecimento da população.A fala emerge como um biomarcador rico e ubíquo, com enorme potencial para o desenvolvimento de ferramentas remotas, de baixo custo e larga escala para o rastreio de várias doenças do foro respiratório e psiquiátrico, bem como doenças neurodegenerativas.

Este potencial deriva do envolvimento dos sistemas respiratório, nervoso e muscular na produção da fala, que, consequentemente, contém informações sobre disfunções em qualquer um desses sistemas. Em todos os níveis da produção de fala, biossinais podem ser captados e estudados para obter informação sobre o seu conteúdo linguístico e paralinguístico.

Esta tese começou por explorar um desses biossinais — os sinais de eletromiografia (EMG) produzidos durante a articulação da fala — contribuindo para estabelecer os fundamentos de uma nova área de investigação, a Paralinguística Computacional Silenciosa. Durante a pandemia COVID-19, o foco foi redirecionado para outros biomarcadores, capazes de ser recolhidos remotamente. Exploraram-se imagens faciais e visual speech como modalidades complementares do sinal acústico para a deteção de apneia obstrutiva do sono, obtendo-se resultados promissores com transfer learning e características baseadas em conhecimento específico, num estudo com 40 sujeitos.

Não obstante os resultados promissores obtidos, persistem dúvidas sobre o que estes modelos realmente aprendem e em que informações se baseiam para fazer a predição de doenças. Assegurar a fiabilidade e a generalização dos resultados é crucial. Esta tese propõe um modelo robusto e interpretável, adequado para a deteção de múltiplas doenças. Propõe-se uma framework que define a fala normativa através de intervalos de referência de características acústicas e linguísticas com significado clínico. Os desvios em relação a esta referência são usados para a classificação de doenças como as de Alzheimer e Parkinson, utilizando diferentes classificadores, nomeadamente Modelos Aditivos Neurais para uma maior interpretabilidade.

Finalmente, numa tentativa de conciliar modelos do tipo caixa negra com interpretabilidade, explora-se o uso de large language models para anotar características de fala de alto nível e baixa dimensionalidade, denominadas macro-descritores (e.g., coerência textual, diversidade lexical). Utilizando apenas quatro macro-descritores, supera-se a deteção da doença de Alzheimer utilizando características convencionais.

Esta tese contribuiu para uma compreensão mais profunda do potencial multifacetado da fala como biomarcador para a saúde, numa perspectiva holística.

Palavras Chave

Fala; Doenças que afetam a fala; Multimorbilidade; Multimodal; Interpretabilidade

Contents

List of Figures

List of Tables

Acronyms

AACD age-associated cognitive decline

AAI acoustic-to-articulatory inversion

ADAS-Cog Alzheimer's Disease Assessment Scale - Cognitive Subscale

ADReSS Alzheimer's Dementia Recognition through Spontaneous Speech (corpus)

AD Alzheimer's disease

AHI Apnea-Hypopnea Index

ASR Automatic Speech Recognition

BERT Bidirectional Encoder Representations from Transformers

BIF bio-inspired features

BLSTM bidirectional long short-term memory

CCC Concordance Correlation Coefficient

CI confidence interval

CLAC Crowdsourced Language Assessment Corpus

CNN convolutional neural network

ComParE Computational Paralinguistics Evaluation

COVID-19 Coronavirus disease 2019

C19C COVID-19 COUGH corpus

CPAP continuous positive airway pressure

CPP cepstral peak prominence

DAIC-WOZ Distress Analysis Interview Corpus – Wizard Of OZ

DL deep learning

DNN deep neural network

DT Decision Tree

ECAPA-TDNN Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural

	Networks
ECoG	Electrocorticography
EEG	Electroencephalography
eGeMAPS	extended Geneva Minimalistic Acoustic Parameter Set
EMA	Electromagnetic Articulography
EMG	Electromyography
FN	false negatives
FP	false positives
GANs	Generative Adversarial Networks
GDPR	General Data Protection Regulation
GMM	Gaussian Mixture Model
HNR	Harmonics-to-noise Ratio
ICD-11	International Classification of Diseases 11th Revision
ILSE	Interdisciplinary Longitudinal Study on Adult Development and Aging
IQR	interquartile range
KB	knowledge-based
LDA	Linear Discriminant Analysis
LIWC	Linguistic Inquire and Word Count
LLM	Large Language Model
LR	Logistic Regression
LRW	Lip Readind in the Wild
MFCC	Mel-frequency cepstral coefficients
ML	machine learning
MMSE	Mini-Mental State Exam
NAM	Neural Additive Model
NN	neural network
OSA	obstructive sleep apnea
PASE	problem-agnostic speech encoder
PC-GITA	Parkinson's disease Corpus from the Applied Telecommunications Group (GITA) at the Universidad de Antioquia, Colombia Parkinson's Spanish corpus
PD	Parkinson's disease
PLDA	Probabilistic Linear Discriminant Analysis

PoS	Part-of-Speech
RBF	Radial Basis Function
ReLU	rectified linear units
RF	Random Forest
RI	Reference Interval
RMSE	root mean squared energy
RT-PCR	reverse transcription polymerase chain reaction
SSD	single shot detector
SVM	Support Vector Machine
TD	time domain
TDNN	time delay neural network
TDNN-F	factorized time delay neural network
TIMIT	Texas Instruments/Massachusetts Institute of Technology corpus
TNR	true negative rate
TN	true negatives
TPR	true positive rates
TP	true positives
TTR	Type-to-token ratio
UAP	Unweighted Average Precision
UAR	Unweighted Average Recall
UF1	Unweighted Average of F1 score
UPDRS	Unified Parkinson's Disease Rating Scale
VAD	Voice Activity Detector
VOT	Voicing Onset Time
WER	Word Error Rate
WOSA	in-the-Wild Obstructive Sleep Apnea Corpus
WSM	in-the-Wild Speech Medical Corpus

1

Introduction

Contents

LESS than half of the world's population is covered by essential health services[1], with access being largely influenced by economic and geographic constraints. "Ensure healthy lives and promote well-being for all at all ages" is the third goal of the United Nations Agenda for 2030, and although until the end of 2019 there were advances in many areas of healthcare, the progress rate was not sufficient and it was further negatively impacted by the Coronavirus disease 2019 (COVID-19) pandemic. One of the many consequences of the COVID-19 health crisis was that people became unable or afraid to seek healthcare services for routine check-ups and timely diagnoses.[2]

It is widely recognized that timely diagnosis, and disease monitoring leads to healthier and longer lives. Recent works [Wurcel et al., 2019] discuss the value of diagnostic information, and how to evaluate it, as they consider that diagnostic information can be a key to guide effective, efficient and affordable health systems.

Recent technological advances, in particular in the machine learning (ML) and deep learning (DL) fields, have allowed the development of highly accurate predictive systems for numerous applications, including medical diagnosis and monitoring tools. Such technologies, allied with the potential of different biomarkers, provide a route for mass screenings of multiple diseases, which would mean a step towards a scalable improvement in global healthcare.

Speech is a rich biomarker that carries information about the speaker's gender[3], age [Bahari et al., 2014; Hechmi et al., 2021], emotions [Schuller et al., 2013b], personality traits [Lee et al., 2021], and temporary states, such as intoxication or sleepiness [Schuller et al., 2011]. Most importantly in the context of this thesis, speech also conveys information on health states, which has motivated many speech-based health applications, including speech therapy [Abad et al., 2013], and the diagnosis and monitoring of speech affecting diseases. The range of these diseases extends beyond the so-called speech and language disorders (e.g. sigmatism, stuttering), and includes neurodegenerative diseases such as Parkinson's disease (PD) [Correia et al., 2021; Pompili et al., 2017], Alzheimer's disease (AD) [Lopez-de Ipina et al., 2018; Zargarbashi and Babaali, 2019], Huntington's disease [Yoon et al., 2006] and Amyotrophic Lateral Sclerosis [Gómez-Vilda et al., 2015]; psychiatric disorders such as depression [Correia et al., 2021], bipolar disease [Karam et al., 2014], and burnout; and diseases that concern respiratory organs, such as obstructive sleep apnea (OSA) [Botelho, 2018; Botelho et al., 2019; Perero-Codosero

[1] Data from: https://sdgs.un.org/goals/goal3 (last accessed 08-02-2021)

[2] https://unstats.un.org/sdgs/report/2020/goal-03/ (last accessed 08-02-2021)

[3] According to the World Health Organization, gender refers to socially constructed roles and norms and sex usually refers to a person's biological characteristics. While the terms are interrelated, they are sometimes conflated or used interchangeably in health data [Kaufman et al., 2023]. This is the case of several speech corpora with gender annotation, and literature on the detection of speech affecting diseases. Thus, in this thesis, the terms *men, women, male, female, gender and sex*, refer to biological sex, encompassing the physical and hormonal characteristics intrinsic to male and female bodies. This differentiation is relevant because physiological differences, such as those affecting the fundamental frequency of speech, manifest differently across sexes. For instance, due to changes in the composition of the vocal folds, in women, pitch tends to decrease with aging and in men, pitch tens to increase with aging. From a health perspective, recognizing the unique speech attributes associated with male and female biological profiles allows for more accurate and sensitive biomarker identification. While this thesis predominantly refers to the binary dimension of biological sex due to the availability of substantial data, it is important to acknowledge that gender is a spectrum and sex may include intersex individuals [Kaufman et al., 2023]. Currently, data on these groups is sparse and therefore not possible to analyse using standard machine learning techniques.

et al., 2019], and COVID-19 [Deshpande and Schuller, 2020]. The references provided describe only a few examples, but many others could be mentioned.

The reason why all these diseases have effects on the speech signal – either more subtle, e.g. in the case of obstructive sleep apnea, or more noticeable, as in the case of a high severity Parkinson's disease – is the fact that speech production is a complex process that involves not only the articulatory muscles but also the nervous and respiratory systems. Speech production starts in the brain with the formulation of the speech intention, which manifests itself as patterns of electrical potentials in the cortex. The electrical signal is conducted through the nervous system to the muscles involved in speech kinematics, and finally, speech is emitted from the mouth as sound waves. At all these speech production levels, biosignals can be captured and studied to retrieve information about the linguistic and paralinguistic[4] content of spoken communication [Schultz et al., 2017].

Other non-invasive biomarkers have been shown to provide valuable information for disease screening, which can be complementary to the speech signal. For instance, cough signals have proven helpful for COVID-19 [Schuller et al., 2021] and tuberculosis detection [Botha et al., 2018], facial images have been used for OSA [Balaei et al., 2017] and depression diagnosis [Zhou et al., 2018], snore signals have been utilized for OSA diagnosis [Calabrese et al., 2009], and text transcripts have been applied for depression [Lopez-Otero et al., 2017] and Alzheimer's disease [Pompili et al., 2020a]. These biomarkers can be collected via web or mobile phone applications, and thus contribute to supporting medical diagnoses during telemedicine appointments, which have proven to be of great importance during pandemic situations, such as the recent one.

Notwithstanding the large potential of the aforementioned biomarkers combined with the strong predictive capabilities of machine learning systems, as well as the encouraging results reported by many researchers, one observes that systems for the automatic detection of speech affecting diseases are not widely integrated into commercial healthcare products [Milling et al., 2022]. Unlike the healthcare sector, many industry sectors, including mobile phones, automobile industry, virtual assistants, have implemented speech and machine-learning based solutions in commercial products. The question that immediately arises is *why?* – Why have these promising and effective technologies not yet been adopted in the healthcare market? Addressing this question requires the consideration of multiple factors, including (but likely not limited to): (i) difficulty of acquiring sufficiently large datasets for health-related tasks which are representative of a population, due to patient-privacy laws, ethical concerns, lack of aware-

[4]Three levels of information can be extracted and studied from the speech signal: linguistic, paralinguistic and extralinguistic information [Laver, 1994]. *Linguistic information* relates to the message being conveyed. At this level, we can extract information such as coherence or lexical diversity of speech, which may be useful for example for the diagnosis of Alzheimer's Disease. *Paralinguistic information* refers to everything that is not linguistic, but is associated with the content of the linguistic message, such as intent, politeness or emotional state. Unlike linguistic and paralinguistic information, *extralinguistic information* is independent of the message being conveyed, as it captures characteristics of the speaker, such as the speaker' habitual voice quality, or pitch range. In this thesis, as is frequent in the speech research community, the term *paralinguistic* is used comprehensively to designate all information which cannot be described only in strictly linguistic terms [Schuller et al., 2013a], thus encompassing both paralinguistic and extralinguistic information. By consolidating these under the term "paralinguistic", we acknowledge the holistic nature of speech as a rich biomarker.

ness in the medical community, and time and money concerns; (ii) the fact that existing datasets and research findings pertain the detection of a single disease compared to healthy controls, when in real-life scenarios where the systems may be deployed, if someone does not suffer from a specific disease, it does not mean they are healthy controls – they may suffer from a different disease whose speech signal is not projected to the same space as healthy controls; (iii) difficulties in interpreting machine learning-based predictions due to the inherent black-box nature of the models; (iv) the fact that machine learning models may capture solely correlations, which may simply reflect biases present in the dataset or other spurious correlations; (v) the fact that some machine learning models are associated with so-called epic failures [McGregor, 2020], which human decision-making processes would certainly prevent.

Most of these reasons are somewhat related to the lack of trust in current ML systems, which sometimes suffer from the problem of shortcut learning, i.e. the learning of decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions, such as real-world scenarios [Geirhos et al., 2020]. Many researchers have argued in favour of more explainable, interpretable, and transparent machine learning systems [Caruana et al., 2015; Stiglic et al., 2020]. Judea Pearl, in "The Book of Why" [Pearl and Mackenzie, 2019], advocates for the importance of transparency, as it enables effective communication. He argues that most limitations of deep learning systems stem from their inability to go beyond what he calls the rung one of the Ladder of Causation: *Association*[5]. This fact may have not hindered some astonishing advances, such as the performance of AlphaGo [Holcomb et al., 2018], but it does hinder learning systems that operate in rich webs of causal forces, such as medicine [Pearl and Mackenzie, 2019]. Pearl argues that ML methods today provide a way to go from finite sample estimates to probability distributions, but we still need to get from distributions to cause-effect relationships [Pearl and Mackenzie, 2019].

Very recently, various subfields of Artificial Intelligence research have been disrupted by the emergence of Large Language Models (LLMs), which have exhibited impressive language generation and understanding capabilities, and have even gained widespread recognition among the general public. Their remarkable performance in numerous tasks has provided the opportunity to reconsider the need for interpretable machine learning – indeed they are able to solve complex problems while providing a human-understandable explanation. This raises intriguing questions: Are these models sufficient for diagnosing language affecting disorders solely based on transcriptions of patients' speech? Alternatively, can their language understanding capabilities be leveraged to better characterize the discourse patterns of patients?

All these aspects, from the different biosignals involved in speech production, the non-invasive biomarkers complementary of acoustic speech, the limitations of current ML-based systems applied

[5]The Ladder of Causation, proposed by Judea Pearl, has three rungs: rung one – *association*, which allows only to see associations between variables, leading to a very superficial understanding of the world; rung two – *intervention*, actors can intervene and change the world, which leads to the understanding of cause-effect relationships; and rung three – *counterfactuals*, where actors can image alternative worlds where the actor would have done something different.

to scarce data, the importance of interpretable models, and the role of the new Large Language Models, define the context in which we frame this thesis, which explores the potential of speech for disease detection. Initially, our focus was on understanding the interconnections of speech production, the different biosignals that precede the sound wave, and the mechanisms through which each disease impacts speech production. This led to the establishment of the novel field of Silent Paralinguistics, i. e., the assessment of speaker states and traits from non-audible spoken communication [Botelho et al., 2020a; Diener et al., 2020]. Our early studies, which retrieve Electromyography (EMG)[6] from the speech signal, and the speaker identity from EMG yielded promising results. Initially, we were aiming at conducting our speech research alongside other modalities such as EMG or Electroencephalography (EEG), however, the generalized COVID-19 lockdown precluded the collection of more data that could provide insights into the manifestations of each disease in the biosignals produced by the nervous systems, stalling this line of research. The pandemic highlighted the importance of delivering timely check-ups and diagnosis remotely, e.g., through telemedicine. Consequently, we shifted our focus to other biomarkers that could be collected non-invasively, unobtrusively, and remotely via web or mobile phone applications. Specifically, we proposed to use visual speech as a new modality for silent paralinguistics and disease detection [Botelho et al., 2021]. We argue that visual speech, also known as lip reading, encodes information on the craniofacial structure, speech articulation, and breathing patterns, which may provide additional relevant diagnostic information. Furthermore, visual speech is captured through face filming, making it suitable for remote acquisition, similarly to acoustic speech.

After exploring the potential of visual speech and facial images to complement the speech signal for disease detection, we discuss the challenges inherent to the datasets used by the research community, which may hinder the reliability of results. A significant issue, which is also reflected in the reasons mentioned earlier that are hampering speech and ML-based applications to support medical diagnosis, is the existence of multiple speech-based datasets tailored for detecting individual diseases. Additionally, most of these datasets are considerably smaller than ideal for the effective application of ML tools, and they are recorded under varying conditions. This is particularly relevant in the context of long longitudinal studies, where recording conditions change over time. Can these datasets be leveraged to build a model that addresses health in a comprehensive manner? We argue that a valid starting point is to characterize reference speech independently of the dataset or recording conditions, and then to use such a definition to identify signature patterns of speech affecting diseases. Particularly, we focus on glass-box predictive models that provide clinically meaningful explanations for the decision process.

Finally, as the new large language models disrupted AI research, we hypothesize that these models can capture high-level clinically-meaningful characteristics of discourse, which can be used for explainable disease screening.

[6] EMG is a technique that captures the electrical signals generated during muscle contraction [Schultz et al., 2017]

1.1 Objectives and research questions

The overall goal of this thesis is *to explore the potential of speech as a trustworthy biomarker to support accessible medical diagnosis*. With this main objective and the previous motivation in mind, we raise more specific research questions that we aim to address. While our current approach to these questions frequently involves the detection of specific diseases through binary classification–comparing each disease to healthy controls–our broader vision is to propose speech as a holistic biomarker for multiple diseases. This marks a step towards scalable remote mass screenings, capturing early signs of various conditions. Although current research is mostly conducted within the scope of individual diseases, our ultimate aim is to move beyond single-disease detection and towards a more comprehensive use of speech in disease screening. The research questions are the following:

RQ1 Do biosignals involved in speech production preceding the sound wave, e.g. EMG, contain paralinguistic information?

RQ2 Are other non-invasive modalities, such as facial images, and visual speech complimentary of speech signals for disease detection? In this context, are in-the-wild data acquired from online repositories a valid alternative to standardized data collection that could provide larger datasets for training ML models?

RQ3 What are the challenges inherent to the existing small datasets for single disease detection? How can we leverage the multiple small datasets to create one generalizable model for health monitoring?

RQ4 Can we determine reference intervals that characterize and define healthy speech, independently of the dataset? Can we leverage this definition of healthy speech to identify signature patterns of each speech affecting disease?

RQ5 How can the advanced text understanding capabilities of new LLMs contribute to the automatic detection of speech-affecting diseases?

1.2 Contributions

The research described through this thesis has been published in ten peer-reviewed articles.

The following two publications summarize our findings concerning the first research question. I was the lead author and contributor in [A] and part of the team in [B].

[A] Botelho, C., Diener, L., Küster, D., Scheck, K., Amiriparian, S., Schuller, B. W., Schultz, T., Abad, A., and Trancoso, I. (2020). Toward silent paralinguistics: Speech-to-EMG – Retrieving articulatory muscle activity from speech. In Interspeech.

[B] Diener, L., Amiriparian, S., Botelho, C., Scheck, K., Küster, D., Trancoso, I., Schuller, B. W., and Schultz, T. (2020). Towards silent paralinguistics: Deriving speaking mode and speaker ID from electromyographic signals. In Interspeech.

Our contributions to the second research question have resulted in the following two publications. Publication [C] focuses on the creation of the in-the-Wild Speech Medical (WSM) Corpus, for Parkinson's Diseases and depression detection. In this publication, my contributions concern the experiments that use i-vectors and x-vectors for classification. I was the lead author in publication [D], which uses a pilot corpus acquired in-the-wild data to perform multimodal detection of obstructive sleep apnea, using speech, facial images and visual speech.

[C] Correia, J., Teixeira, F., Botelho, C., Trancoso, I., and Raj, B. (2021). The in-the-wild speech medical corpus. In ICASSP. IEEE.

[D] Botelho, C., Abad, A., Schultz, T., and Trancoso, I. (2021). Visual speech for obstructive sleep apnea detection. In Interspeech.

Our findings regarding the third research question were summarized in the following publications. In publication [E], we explore different methods to detect COVID-19 from cough signals. My contributions concern the experiments using TDNN-F embeddings. The relationship to this research question is to highlight possible unexpected biases in the datasets — in this case a bias introduced by different bandwidth of the recordings of diseased and control subjects. In publication [F] we describe a parallel analysis of two corpora: a longitudinal conversational corpus in German, and a cross-section English corpus containing image descriptions, for the automatic detection of Alzheimer's disease. We observe that with different corpora, different features seem to be more relevant for Alzheimer's disease detection, which raises questions on transferability of results across small datasets. In this publication, Ayimnisagul Ablimit and I were joint lead authors. Motivated by the results published in [F], we explored the challenges of using cross-domain corpora or longitudinal corpora in the context of pathological speech studies, which were published in [G], where I was also lead author.

[E] Solera-Ureña, R., Botelho, C., Teixeira, F., Rolland, T., Abad, A., and Trancoso, I. (2021). Transfer learning-based cough representations for automatic detection of covid-19. In Interspeech.

[F] Ablimit, A.*, Botelho, C.*, Abad, A., Schultz, T., and Trancoso, I. (2022) Exploring dementia detection from speech: cross corpus analysis. In ICASSP. (*equal contribution of first two authors)

[G] Botelho, C., Schultz, T., Abad, A., and Trancoso, I. (2022). Challenges of using longitudinal and cross-domain corpora on studies of pathological speech. In Interspeech.

Our work addressing the fourth research question has led to the publication of one conference paper [H] and one journal paper [I], where I served as the lead author and contributor. In publication [H], we explore the idea of characterizing reference speech, using reference intervals, a concept traditionally used in clinical laboratory science, which we adapt for speech analysis in the context of health. Publication [I] expands upon this concept by proposing to leverage the definition of reference speech for the detection of multiple speech-affecting diseases. Our approach applies Neural Additive Models, a class of transparent neural networks, to enhance the interpretability of our findings.

[H] Botelho, C., Abad, A., Schultz, T., and Trancoso, I. (2023). Towards reference speech characterization for health applications. In Interspeech.

[I] Botelho, C., Schultz, T., Abad, A., and Trancoso, I. (2024). Speech as a Biomarker for Disease Detection. – *submitted to IEEE Access*.

Finally, publication [J] addresses our fifth research question by exploring the ability of LLMs to detect Alzheimer's Disease from picture description transcriptions and to annotate macro-descriptors, which are high-level dimensions useful for disease detection. I was lead author and contributor in this publication.

[J] Botelho, C., Mendonça, J., Pompili, A., Schultz, T., Abad, A., and Trancoso, I. (2024). Macro-descriptors for Alzheimer's disease detection using large language models. In Interspeech.

Besides the peer-reviewed publications, I gave an expert talk at ICASSP 2022 – Singapore, together with Ayimnisagul Ablimit. The talk, entitled "Speech as a disease biomarker", summarizes our work, while discussing the challenges in the field.

1.3 Thesis structure

The remainder of this thesis begins with a discussion of the fundamental concepts necessary for understanding the associations between speech alterations and the various speech affecting diseases, in chapter 2. This chapter starts by explaining the speech production process and proceeds to describe various speech affecting diseases, detailing their impact on speech. It further discusses how these diseases are often risk factors for each other and how their effects on the speech signal partially overlap. Consequently, it provides our perspective, advocating for a global view on health and the use of speech as a biomarker for multiple diseases. This perspective is particularly relevant for guiding the rest of this thesis.

Chapter 3 introduces the health-related speech corpora available to the research community, as well the methods for the automatic detection of speech affecting diseases, including features, evaluation metrics and strategies to cope with limited size dataset.

Chapter 4 summarizes our exploratory work on the novel field of silent computational paralinguistics, and the interconnection between speech and another biosignal involved in the speech production process – EMG. In particular, the chapter focuses on using speech signals to retrieve the underlying EMG signals. This type of speech-to-EMG conversion facilitates a deeper understanding of the interconnections among various biosignals and their roles in communication. Additionally, it holds potential applications in medicine, particularly in enhancing articulatory awareness during speech therapy. Further applications include computer animation, enabling visualization of realistic muscle movements [Sagar and Scott, 2009], or as a means to generate large amounts of synthetic EMG data from audio, decreasing the amount of costly laboratory recordings of EMGs.

Chapter 5 concerns the multimodal automatic detection of obstructive sleep apnea, a sleep-related breathing disorder which is estimated to affect over one in every eight individuals worldwide, most of which are currently undiagnosed. We focus on biosignals that can be collected remotely, e.g. through a telemedicine appointment: speech, facial images and visual speech – a modality which we introduced, for the first time, for paralinguistic applications and that builds on our previous finding that muscle contraction captured by EMG signals on the face encode paralinguistic information. The work for OSA detection uses in-the-wild data as an alternative for standard medical data collection. In this chapter, we further discuss the suitability of using such large datasets available online in multimodal repositories for disease detection.

Although the results presented in chapter 5, as well as other works in the literature regarding the detection of speech affecting diseases, are very promising, there are a few issues to address before a transition to commercial products can take place. Such issues, which mostly relate to the robustness and generalizability of results to new datasets, are explored in chapter 6. The chapter presents a collection of experiments that highlight some of the challenges introduced by the datasets used by the speech community for disease detection. Concretely, it discusses unexpected biases; the difficulty of translating models and results to new domains, including different recording conditions, speech tasks and languages; and finally it explores how much information about the recording conditions is encoded in the features that are typically used to detect diseases from speech.

These challenges motivated the development of an interpretable approach suitable for the detection of multiple diseases. Chapter 7 presents a framework that first defines reference speech, and then leverages this definition to perform disease detection. Reference speech is characterized using reference intervals for clinically meaningful acoustic and linguistic features derived from a reference population, a novel approach in the speech health field inspired by biochemistry's use of reference intervals for medical diagnostics. We then quantify deviations of new speakers from this reference model, and use these deviations as input to detect Alzheimer's and Parkinson's disease. One classification strategy explored is based on Neural Additive Models, a type of glass-box neural network. The chapter also discusses

strategies for multidisease classification, contingent on the availability of adequate data.

Chapter 8 discusses the role of LLMs in the detection of language-affecting diseases, under two scenarios. In the first scenario, we evaluate LLMs' performance when directly queried for disease prediction. In the second scenario, we explore their potential as annotators of high-level discourse characteristics, designated macro-descriptors, which are then used for detecting Alzheimer's disease. The experiments described compare the performance of both open access and closed source LLMs, based on manual and automatic transcriptions, and using several prompting strategies. We also explore whether pause information is able to further aid Alzheimer's disease detection.

Chapter 9 summarizes and discusses the main findings of each chapter and outlines directions for future research. The chapter concludes with final remarks, briefly addressing ethical considerations and highlighting the importance of explainability and reliability in the automatic detection of speech-affecting diseases.

2

Speech affecting diseases

Contents

THE fact that speech production is a complex process that involves the respiratory, nervous, and muscular systems implies that disruptions in any of these systems can perturb the speech signal. Consequently, speech can encode information indicative of diseases affecting these systems. This chapter begins with an overview of the speech production process[1] to facilitate reasoning on which speech alterations are associated with each disease studied in this dissertation, and how to map these alterations to measurable features or predictable outcomes. More comprehensive descriptions exist in the literature, e.g. refer to [Singh, 2019]. Afterwards, we characterize common speech affecting diseases, including OSA, COVID-19, AD, PD, and depression, and discuss their impacts on the speech signal.

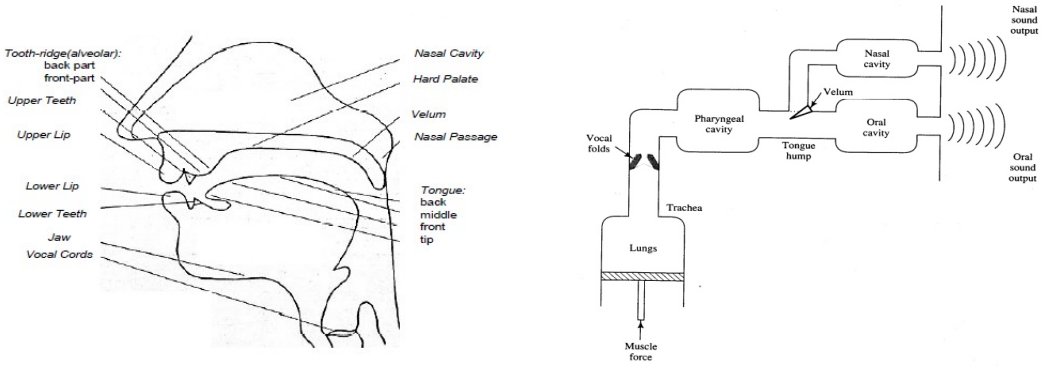
Given the association of several of these diseases with aging, we also include sections on aging and multimorbidity. This chapter concludes with our perspective on the importance of a global view of health. This is particularly relevant because speech-affecting diseases are often considered risk factors for each other, and their effects on speech frequently overlap. This perspective is integrated into an article published at the IEEE Access Journal Botelho et al. [2024].

2.1 Speech production

The speech production process starts in the cerebral cortex, with input from several other areas in the brain, including those of artistic expression, cognitive thought, planning, and hearing. The information is then transmitted through the nervous system to allow a coordinated activity of the muscles in the larynx, chest and abdominal regions, the vocal folds, and all the articulators [Singh, 2019]. The lungs expel the air, which passes through the trachea and the larynx, where the vocal folds are housed. At the vocal folds, there are some pressure differentials which cause them to open and close cyclically, resulting in the release of hundreds of air pulses per second, generating sound waves. This process is called *phonation*. The sound waves produced during phonation pass through the vocal tract, schematized in Figure 2.1. The vocal tract is composed of resonance chambers and articulators, which filter the excitation signal to produce distinct sounds that compose the languages we speak. Simultaneously, a feedback mechanism is activated from the ears, and from visual input, if present [Singh, 2019]. These systems and processes are described with more detail in the sections below.

Usually, all these components function in close coordination, but it is not always the case. For instance, we can have audible speech without phonation, which is the case of whisper. We can also have phonation without oral articulation as in some aspects of yodeling that depend on pharyngeal and laryngeal changes; and we can have silent speech without breath and voice, and thus without acoustic output, e.g. for lip reading. Furthermore, there are also two other types of speech without

[1] Most of this section uses the description of speech production by Encyclopedia Britannica [Arnold et al., 2019] as a reference. In the absence of an explicit reference, [Arnold et al., 2019] should be considered.



(a) **(b)**

Figure 2.1: (a) Vocal tract and the main articulators involved in speech production [Huang et al., 2001]. (b) Block diagram of human speech production.

acoustic output, which encompasses only the cognitive formulation of speech: imagined speech and inner speech. Imagined speech is similar to silent speech, but without the actual movements of the articulators. Imagined speech corresponds to a motor imagery of speaking, in which the speakers should feel as though they are producing speech rather than simply talking to themselves. Given that imagined speech is produced without any articulatory movements, this speaking mode requires observations at the neural level [Schultz et al., 2017]. Inner speech corresponds to an internalized process in which one thinks in pure meanings [Schultz et al., 2017].

The nervous system

The nervous system controls the speech production process, both at a conscious and unconscious level. The human brain possesses several language centers in the cortex, including the Brocca's area, the Wernicke's area, and the cortical hearing center, as represented in Figure 2.2. These language centers are interconnected with subcortical areas, such as the thalamus, responsible for emotional integration, and the cerebelum, responsible for movement coordination.

Brocca's area is located in the frontal lobe of the brain cortex, and is responsible for the motor control of the movements for expressive language. From the cells in Brocca's area, fibers emerge that eventually connect with the cranial and spinal nerves that control the muscles of oral speech. Injury in this region causes expressive aphasia, i.e., the inability to speak or write. The *Wernicke's area*, located in the temporal lobe of the cortex, is responsible for receptive speech comprehension. An injury in this area results in receptive aphasia, i.e., inability to understand spoken or written communications, as if the person had never learned the language before. The *auditory cortex* receives and processes the contents of sounds, voices, or music.

Speech production is further regulated by feedback mechanisms – the auditory feedback and the proprioceptive sense. The auditory feedback mechanisms, enabled by the auditory cortex which receives and processes the contents of sounds [Sitek et al., 2013], informs the speaker about pitch, volume,

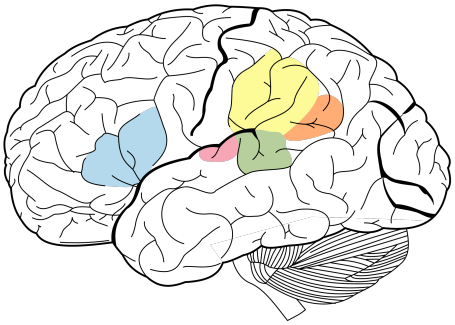


Figure 2.2: Language centers in the brain. The Angular Gyrus is represented in orange, Supramarginal Gyrus is represented in yellow, Broca's area is represented in blue, Wernicke's area is represented in green and the Primary Auditory Cortex is represented in pink.[a]

[a]Author credits: By James.mcd.nz - self-made - reproduction of combined images Surfacegyri.JPG by Reid Offringa and Ventral-dorsal streams.svg by Selket, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=3226132

inflection of the voice, articulation precision, selection of appropriate words, and other audible features. The proprioceptive sense provides continuous information on the position of the muscles, tendons, joints, and other moving parts. Limitations of these systems compromise the quality of speech, as observed in pathological scenarios, such as deafness, paralysis, and underdevelopment.

Examples of diseases that may condition the normal functioning of the nervous systems are Alzheimer's disease and Parkinson's disease.

Laryngeal system

Voice originates in the larynx. Structurally, the larynx is composed of the vocal folds and a frame of three cartilages, as shown in Figure 2.3. Most of these cartilages ossify to variable degrees at different ages, in the presence of masculinizing hormones. The cartilages and the muscles that are attached to them can adjust the shape, position and mechanical motion of the vocal folds [Singh, 2019]. The muscles of the larynx can be subdivided into extrinsic (thyropharyngeus, cricopharyngeus, and the short cricothyroid[2]), which move the larynx as a whole, and intrinsic, which move the vocal folds to shape the glottis. The larynx is enervated by two nerves, the superior laryngeal nerve and the recurrent or inferior laryngeal nerve. The shape of the larynx is compared to the shape of a horn. A well-shaped, wide and flexible larynx enhances the projective potential of the voice. In turn, a morphologically narrow, pathologically constricted, or emotionally tightened larynx leads to a constricted voice sound with poor carrying power.

The vocal folds or vocal cords are two membranes held together that oscillate to produce voiced sounds. These oscillations are self-sustained vibrations maintained by pressure differentials. During expiration, the air in the lungs is driven upwards through the trachea against the undersurface of the vocal folds. This increases the subglotic pressure until it exceeds the closing effort of the vocal cords. When that happens, a puff of air escapes, lowering the subglotic pressure, and closing again the vocal

[2]Although the short cricothyroid muscle is actually external to the larynx, it is typically discussed within the intrinsic muscles.

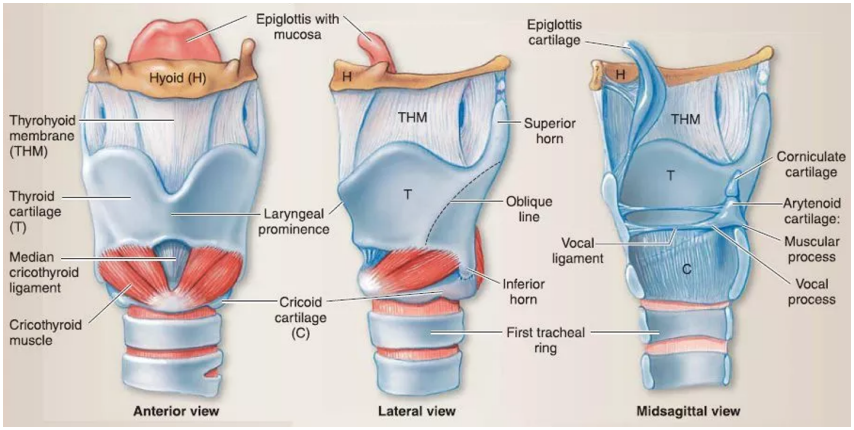


Figure 2.3: Larynx anatomy, from [Harrel and Dudek, 2019].

cords. Then, the subglotic pressure raises again and the cycle is repeated. The frequency of this cycle corresponds to the fundamental frequency of the sound wave generated, *F0*, and the pitch we perceive in the speech sound. When the vocal folds are too tense and do not vibrate periodically, the sound emitted is unvoiced.

The vocal folds are a layered structure, with three layers: an inner muscular layer, a middle soft tissue layer called the lamina propria, and an external epithelial layer. The lamina propria is a layer of approximately 1 mm composed of an extracelular matrix, mainly constituted of the proteins collagen and elastin that are responsible for the elastic properties of the vocal folds [Singh, 2019]. The fibers of collagen and elastin are thought to play an important role in glottal closure, *i.e.*, in the closing of the vocal folds during phonation.

There are many biological parameters that affect phonation and thus the characteristics of the voice signal. Some of these parameters are stable for a given individual across their lifetime, and others are variable for the same individual. Particularly, parameters that affect F0 include the size of the larynx, which in turn, is conditioned by the body type, sex, and age. The size of the larynx is approximately the same in small boys and girls, while an adult male larynx is approximately three times the size of a female larynx. Shorter vocal chords produce higher-pitched voices. The fundamental frequency of the speech sound, F0, is typically around 100 Hz for male and 200 Hz for female speakers. Within the individual fixed range, there are other variables impacting the pitch, which include the tension of the cords, the force of the glottal closure, and the expiratory air pressure. The increased tension of the cricothyroid muscles increases the pitch, and vice-versa [Arnold et al., 2019]. Under pathological conditions, for instance, the glottal closure may not be complete, which means the energy in the higher harmonics is reduced and the voice is perceived as “weak” [Singh, 2019]. Also, as people age, there is a decrease of collagen and elastin in the body (thus, also in the lamina propria of the vocal folds), which causes the vocal folds to be unable to close completely with the negative intraglottal pressure, which again leads

to a “weak voice” [Singh, 2019]. The hormonal state of the speaker has also been associated with F0 variations [Singh, 2019] and voice quality: for instance, when a female’s voice is lower or rougher, it may indicate the presence of virilizing hormones, which cause excessive ossification of the larynx cartilages. In reverse, the lack of these hormones in males leads to a deficit in ossification, resulting in a voice that sounds more female-like.

Examples of diseases that may impact the laryngeal system are obstructive sleep apnea, and COVID-19, as will be further discussed later.

Articulatory system

The articulatory system includes three cavities – the nasal cavity, the oral cavity, and the pahryngeal cavity – which are the resonators; and the articulators, such as the soft and hard palate, the tongue, teeth, and lips. For any sound, depending on the configuration of the vocal tract, resonances occur at specific frequencies, or at a narrow band of frequencies around a central frequency, due to the soft tissue. These resonant frequencies are called *formants*. The identity of each sound unit is largely coded in the formats, which are a consequence of the configuration of the articulators.

The soft palate, or velum, acts as a valve enabling the passage of air through the nasal tract. When the soft palate isolates the oral cavity from the nasal cavity, oral sounds are produced; when the separation is not complete, nasal sounds are produced instead. The hard palate enables the production of consonants when the tongue is placed against it, and when the tongue is placed away from the palate, vowels are produced, which makes the tongue an articulator organ. The teeth are also responsible for articulation, enabling the tongue to be placed against it to produce certain consonants. The lips can be shaped differently to affect vowel quality, or closed to stop the airflow to produce certain consonants, such as /p/, /m/, and /b/ [Huang et al., 2001]. The configuration of these supraglottic structures is very characteristic of each individual, which makes each person’s voice unique, and as identifiable as fingerprints are [Arnold et al., 2019]. The anatomic shape and physiologic flexibility of the supraglottic structures are impacted by the inborn shape, and the learned behaviour of using it for communication. For instance, a native speaker of a given language learns certain articulatory patterns which remain audible in all other languages learnt after puberty [Arnold et al., 2019].

An example of a disease that may impact the articulatory system is Parkinson’s disease, as will be described later in this chapter.

2.2 Common speech affecting diseases

Speech affecting diseases can be grouped into four broad categories: speech and language disorders such as sigmatism, and stuttering; neurodegenerative diseases, such as PD, AD, Huntington’s disease and Amyotrophic Lateral Sclerosis; psychiatric disorders such as Depression, Bipolar disease, and

burnout; and diseases that concern respiratory organs, such as OSA, tuberculosis, COVID-19, and the common cold. These four categories do not constitute a formal classification. For an official categorization of these diseases according to the International Classification of Diseases 11th Revision (ICD-11), refer to Appendix A. In this work, we will address the automatic detection of OSA, COVID-19, AD, PD, and depression. These diseases have a high worldwide prevalence and constitute examples of respiratory, neurodegenerative and psychiatric disorders. This section describes each of these diseases and their implications in the speech signal.

Another relevant categorization for characterizing how speech is affected by different diseases is the one used by Monoson and Fox [1987], who distinguish three types of speech anomalies: *Resonance anomalies* characterized by an abnormality in voice quality, which can be associated with an abnormal coupling of the vocal tract and oral tract, or related to an abnormality in vocal tract damping and compliance – "cul-de-sac" resonance; *Articulatory anomalies* characterized by a mispronunciation of speech sounds, associated with faulty function or control of the articulators (tongue, lips, hard and soft palate); and *Voice anomalies*, later also called *phonation anomalies* result from a laryngeal dysfunction.

2.2.1 Obstructive Sleep Apnea (OSA)

Obstructive Sleep Apnea (OSA) is a sleep-concerned breathing disorder characterized by repetitive pharyngeal collapse during sleep, which results in complete stops or decreases of the breathing airflow, despite continued or increased inspiratory efforts [Arnold et al., 2017]. This leads to decreased blood oxygen levels and increased carbon dioxide levels. In response to the increased inspiratory effort associated with hypoxia and hypercania, the patients arouse from sleep with activation of the sympathetic nervous system, which causes sleep fragmentation, neurocognitive sequelae and cardiovascular sequelae [Malhotra and White, 2002; Ong and Crawford, 2013]. Patients with OSA report decreases in their quality of life, mood and personality changes, relationship discord associated with loud snoring [Paiva et al., 2014], depression, cognitive impairment, and excessive daytime sleepiness [Punjabi, 2008]. OSA is also associated with diabetes, hypertension and cardiovascular diseases [Arnold et al., 2017; Pozo et al., 2009]. Furthermore, there is growing evidence of a bidirectional relationship between sleep disorders and dementia [de Chazal et al., 2020].

It is estimated that approximately one thousand million adults aged 30 to 69 years old have OSA [Benjafield et al., 2019], which represents one-seventh of the world's adult population [Lyons et al., 2020]. These numbers tend to increase with the growth of OSA's main risk factors: obesity and population aging. Untreated OSA imposes extensive costs for the individuals, their families, and the society at large regarding the economy, health system, and public safety [Armeni et al., 2019; Lyons et al., 2020; Sullivan, 2016].

The severity of OSA is quantified using the Apnea-Hypopnea Index (AHI), which measures the num-

ber of apneas and hypopneas per hour. OSA diagnosis requires an AHI of more than 5 respiratory events per hour with maintained or increased inspiratory effort, associated with typical symptoms: daytime fatigue or sleepiness, unrefreshing sleep, loud snoring, and witnessed breathing interruptions by the bed partner. In the absence of the typical symptoms, the diagnosis requires a minimum of 15 respiratory events per hour [Paiva et al., 2014]. The gold standard diagnosis of OSA is based on polysomnography, a medical exam that is time consuming, expensive, and uncomfortable for the patient and the family [Kriboy et al., 2014b]. Moreover, this method is unable to keep up with the growing number of cases and thus unlikely to meet future demands [de Chazal et al., 2020]. Consequently, more comfortable procedures are needed that are both cost-effective and less time-consuming.

Effects on speech

The pathophysiological mechanisms responsible for OSA, described in the literature, e.g. in [Malhotra and White, 2002], have an impact on the speech production process, not only through the compromised pharyngeal anatomy characteristic of OSA patients, but also associated neuromotor dysfunction. The compromised pharyngeal anatomy characteristic of apnea patients can be explained by certain craniofacial features, including skeletal and/or soft tissue components [Sutherland et al., 2012]. Skeleton components include a shorter mandible corpus, smaller mandibular enclosure area, retrognathia of the mandible, maxillary constriction and shorter length, narrow cranial base, inferiorly positioned hyoid bone, longer anterior face and extended head position. Soft tissue components include enlarged tongue, uvula and soft palate, larger lateral pharyngeal wall and parapharyngeal fat pad volumes, smaller upper airway space and imbalance between tongue size and craniofacial enclosure [Lyons et al., 2020; Sutherland et al., 2012].

According to Monoson and Fox [1987], the speech of OSA patients is characterized by three types of speech anomalies described earlier: resonance, articulatory and phonation anomalies. These three types of anomalies can result from altered structure or function of the vocal tract, above described as characteristic of OSA. In their study in 1987, where trained judges listened to the speech of 13 subjects suffering from sleep apnea, Monoson and Fox [1987] concluded that (quoting) “the majority of sleep apnea individuals we have encountered can be judged to have a speech disorder apparently related to vocal tract dysfunction resulting in altered articulation and resonance.” In a later study by the same authors [Fox et al., 1989], judges listened to the speech of 27 sleep apnea subjects. The judges were able to identify differences between sleep apnea patients and controls in terms of articulation, resonance and phonation. Surprisingly, phonation was found to be the most relevant discriminator between sleep apnea and controls. Pozo et al. [2009] associates the phonation anomalies to the heavy snoring of OSA patients, which can cause inflammation in the upper respiratory system and affect the vocal cords. Furthermore, Pozo et al. [2009] have also found significant differences between OSA and control group speakers in terms of relative levels of nasalization between different linguistic contexts.

2.2.2 Coronavirus disease 2019 (COVID-19)

The COVID-19 respiratory disease is an infectious disease caused by the SARS-CoV-2 virus [WHO, 2022a]. It was declared a pandemic by the World Health Organization in 11 March 2020 and has had dramatic personal, societal and economical consequences that extend until today.

The main symptoms of COVID-19 include fever, coughing, and breathing difficulties. Other symptoms include chills, repeated shaking with chills, muscle pain, headache, sore throat, and loss of taste or olfact. Most people infected with the virus experience mild to moderate respiratory illness and recover without requiring special treatment, or are completely asymptomatic. However, some people become seriously ill and require medical attention. People at risk of developing more serious symptoms include older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer [WHO, 2022a].

Clinical diagnosis of COVID-19 relies on reverse transcription polymerase chain reaction (RT-PCR) and antigen tests. However, these procedures present several disadvantages: significant monetary cost, intrusive and in-person collections of samples performed by professionals that strain public health systems, diagnosis delays due to saturation of laboratories, etc. For all these reasons, we observed an increasing interest in developing reliable, cost-effective, immediate and easy to use tools that can help health care operators, institutions, companies, etc. to optimize their screening campaigns.

Effects on speech

COVID-19 is a disease that affects primarily the respiratory tract [Shah et al., 2022]. Reports indicate that between 60% and 80% of symptomatic patients develop a dry cough, and roughly a third develop a wet cough [Song et al., 2021]. Patients also report fluid accumulation in the lungs and respiratory distress [Suess and Hausmann, 2020]. These respiratory impairments naturally impact patient vocalizations, influencing both coughing and voice quality [Shah et al., 2022]. Asiaee et al. [2020] have posited that recurrent coughing may lead to inflammation or degeneration of vocal fold tissue, potentially increasing jitter and shimmer in patients with COVID-19. This vocal fold trauma could further result in air leakage and incomplete closure of the vocal folds, conditions associated with a breathier voice and increased spectral noise, which in turn results in decreased Harmonics-to-noise Ratio (HNR) and cepstral peak prominence (CPP). Asiaee et al. [2020] also report a decreased maximum phonation time, which is indicative of airflow insufficiency due to reduced lung volume. According to [Singh, 2019], a person's lung capacity is associated with the intensity and duration of the intervocalic breath, the duration of inter-breath intervals, as well as the amplitude of the signal during standardized normal conditions, and total utterance duration. However, these parameters may also be affected by other factors, such as speaking style, emotions and mental states.

Other studies have also identified that the motion of the vocal folds is adversely affected in symptomatic COVID-19 patients [Al Ismail et al., 2021; Deshmukh et al., 2021]. In particular, by visual com-

parisons between oscillation patters of healthy and symptomatic COVID-19 patients, anomalies can be observed, such as asynchrony, motion asymmetry, reduced oscillation range, etc. [Al Ismail et al., 2021].

In a self-assessment study, COVID-19 patients reported difficulty producing certain voiced sounds and noticed changes in their voice [Lechien et al., 2020].

2.2.3 Alzheimer's disease (AD)

Fifty-five million people worldwide live with dementia – a syndrome that leads to the gradual decrease in multiple cognitive functions beyond what is considered normal in biological aging [WHO, 2021 (accessed on September 28, 2021]. The most frequent form of dementia, Alzheimer's disease (AD), is a progressive neurodegenerative disorder characterized by loss of neurons and synapses in the cerebral cortex and in certain subcortical regions. At an early stage, AD causes impairments in memory, language, and spatio-temporal orientation. As the disease progresses, other alterations arise, including visuo-spatial deficits, changes in abstraction and judgment, and later apraxia (difficulty in organizing motor actions intentionally).

AD is typically diagnosed when there are cognitive or behavioral symptoms that interfere with the person's ability to perform usual daily activities or work [McKhann et al., 2011]. This diagnosis is subjective, and is often confused with normal biological aging or stress. However, detailed neuropsychological testing is possible, and can reveal mild cognitive difficulties up to eight years before a person fulfills the clinical criteria for the diagnosis of AD [Bäckman et al., 2004]. Examples of these detailed neuropsychological tests are the Mini-Mental State Exam (MMSE), and the Alzheimer's Disease Assessment Scale - Cognitive Subscale (ADAS-Cog) [Rosen et al., 1984].

Effects on speech

Although memory impairment is the most prominent symptom of AD, language impairments are also prevalent. In particular, the speech of patients with AD is characterized by word-finding difficulties, repetitions, reduced vocabulary, an overuse of indefinite and vague terms, and inappropriate use of pronouns [Forbes et al., 2002; Oppenheim, 1994]. Furthermore, the discourse of AD patients is described as fluent but not informative, characterized by incomplete and short sentences, and lacking coherence and cohesion [Hier et al., 1985; Pompili, 2019]. Several authors have also encountered alterations in temporal parameters of speech, such as hesitation rate and speech tempo, that may indicate deficits in underlying cognitive processes, such as speech planning, structural organization, and production [Hoffmann et al., 2010].

More detailed characterizations of AD and its impacts in speech can be found, for example, in [Pompili, 2019] and [de la Fuente García, 2021]. Voleti et al. [2019] provides a very complete review of existing speech and language features for studying cognitive and thought disorders, including AD. Boschi et al.

[2017], and Hecker et al. [2022] also provide exhaustive reviews on voice analysis for recognizing neurological disorders, including AD.

2.2.4 Parkinson's disease (PD)

Parkison's disease (PD) is the second most common neurodegenerative disorder in the western world, after AD and the most common movement disorder, affecting about 1% of people over 60 years old [De Lau and Breteler, 2006]. PD is characterized by degeneration of dopaminergic neurons in the brain, resulting in dopamine deficiency [Dallé and Mabandla, 2018]. Although the cause of the degeneration is not established, Dallé and Mabandla [2018] summarizes mechanisms that may contribute to the vulnerability of the dopaminergic neurons. PD is mainly considered an idiopathic disease, i.e., a disease with uncertain origin apparently arising spontaneously, but studies have reported that it can also have genetic or environmental origins, and both origins can be associated with early life stress [Dallé and Mabandla, 2018].

The diagnosis of PD is generally made when patients are around sixty or seventy years old, although some cases are found in people in their forties. The standard method to diagnose and evaluate the neurological state of Parkinson's patients is through the revised version of the Unified Parkinson's Disease Rating Scale (UPDRS), provided by the Movement Disorders Society [on Rating Scales for Parkinson's Disease, 2003].

The major motor symptoms of PD include resting tremor (shaking of a body part when at rest), rigidity (resistance to movement when trying to move), akinesia (absence of normal unconscious movements), bradykinesia (slowness of movement), hyperkinesia (reduction in movement amplitude), and postural instability (impaired balance of the body). These motor symptoms are often accompanied by non-motor symptoms including anxiety, depression and/or impairment in cognitive functions [Dallé and Mabandla, 2018]. These motor symptoms are secondary to the neuronal degeneration that occurs in the central nervous system, several years prior to the onset of the clinical symptoms – a preclinical silent phase that often includes depression symptoms [Dallé and Mabandla, 2018]. A critical analysis and deeper research on this preclinical asymptomatic phase may help monitor people at risk of developing the disease, and predict it before the onset of the symptoms.

Effects on speech

About 89% of PD patients develop speech impairments [Ramig et al., 2008], most commonly hypokinetic dysarthria. This condition is characterized by weakness, paralysis, and lack of coordination in the motor speech system, affecting respiration, phonation, articulation, and prosody. Deficits in phonation are caused by inadequate closing of the vocal folds which is associated with muscle rigidity, and result in a breathy or hoarse voice quality. Ma et al. [2020] describes decreased HNR in PD patients, associated with an asthenic or "weak voice"; increased jitter, which indicates unstable vocal fold vibration and is

associated with a rough vocal quality; and increased shimmer, associated with breathiness. Phonation problems are typically the first to occur, and as the disease progresses the other impairments gradually appear [Ramig et al., 2008].

Articulatory impairments are associated with reduced amplitude and velocity of the movements of the lips, jaw, and tongue. In particular, patients may have difficulties pronouncing stop consonants, such as /p/, /t/, /k/, /b/, /d/, /g/, and repetitions of consonant-vowel combinations.

Prosodic impairments are evident on suprasegmental analysis, and include variations in intonation (monopitch), and loudness (monoloudness) [Ramig et al., 2008].

Overall, these deficits contribute to reduced speech intelligibility and naturalness. Detailed characterizations of speech produced by PD patients can be found in [Arias-Vergara, 2022; Pompili, 2019].

It is also important to note that the medication frequently taken by Parkinson's disease patients, as well as by patients with other diseases, may also have an impact on vocal characteristics [Goberman et al., 2002; Ma et al., 2020; Pompili et al., 2020c].

2.2.5 Depression

Depression, also referred to as major depressive disorder or clinical depression, is a common mental disorder, affecting 5% of adults worldwide [WHO, 2022b]. Late-life depression, i.e., depression that affects adults older than 65 who have never had depressive symptoms before, is also very common. Depression is characterized by persistent feelings of sadness and hopelessness, as well as loss of interest or pleasure in activities previously enjoyed [WHO, 2022b]. Serious cases of depression can lead to suicide, which, in turn, is responsible for the death of 700,000 people every year [WHO, 2022b]. There are interrelationships between depression and physical health. For example, cardiovascular disease can lead to depression and vice versa. These interrelationships are further described in section 2.3.

Despite the high prevalence and enormous societal impact of depression, its diagnosis remains subjective and heavily dependent on patients' cooperation and physicians' expertise [Cummins et al., 2011]. In countries of all income levels, people who experience depression are often not correctly diagnosed, and others who do not have the disorder are too often misdiagnosed and prescribed antidepressants [WHO, 2022b]. The standard approach for diagnosing depression, according to the American Diagnostic and Statistical Manual of Mental Disorders, 5^{th} Edition [Association et al., 2019], is to verify that the individual being diagnosed experiences depressed mood or loss of interest and pleasure, and four or more symptoms out of a list, which includes weight gain or loss when not dieting; increased or decreased appetite; slowing down of thought and a reduction of physical movement (observable by others); sleep disturbance; psychomotor agitation or retardation; feelings of worthlessness or guilt; diminished ability to think or concentrate, or indecisiveness; and recurrent thoughts of death or suicidal ideation. These symptoms must coexist during the same two-week period [Cummins et al., 2015b] for confirming the

diagnosis. The criteria-based diagnosis of depression can also be accomplished with other instruments, including several self-administered questionnaires, such as the Patient Health Questionnaire [Kroenke and Spitzer, 2002].

Effects on speech

Besides the symptoms already described, the speech of depressed subjects is also altered. Concretely, it is characterized as dull, monotonous, lifeless, quieter and with reduced loudness variations [Cummins et al., 2015b; Singh, 2019]. The psychomotor retardation characteristic of depression leads to a tightening of the vocal tract, and consequently phonation anomalies that are expected to result in smaller range of formant frequencies [Flint et al., 1993]. The impairments in the control of glottal mechanisms are also expected to increase jitter and shimmer [Kliper et al., 2015]. The psychomotor retardation may also be associated with a decreased speaking or articulation effort. Pause frequency and duration are also expected to be altered. Cummins et al. [2022] found that depressed subjects speak slower than controls, regardless of the language spoken, when comparing speakers of English, Dutch and Spanish. Singh [2019] describes changes in the breathing patterns, with silences occurring in inappropriate locations considering the linguistic content, reduced syllabic stress and harsher and breathier voice quality.

Besides these acoustic queues, the speech of depressed subjects is also marked by linguistic features. Rude et al. [2004] found that depressed individuals use more first-person singular pronouns, more negative emotion words and less positive emotion words. Although different authors have found contradictory results, Tølbøll [2019] has conducted a review of 26 papers, and found correlations between these three markers and depression. The usage of first-person singular pronouns has been associated with the hypothesis that depressed individuals are in a high self-focused state, which builds on the theory of the psychologists Jeff Greenberg and Tom Pyszczynski [Pyszczynski and Greenberg, 1987], that depressed subjects become trapped in a bad self-regulatory cycle [Tølbøll, 2019]. The usage of more negative emotion words has been explained by Beck's cognitive theory of depression [Beck, 1967], which states that depressed individuals experience themselves, others, and the future in a negative way [Tølbøll, 2019].

For a very complete description of how depression impacts speech, refer to [Cummins et al., 2015a].

2.3 Multimorbidity and geriatric health

By 2050, the global population over 60 years old is estimated to reach 2.1 billion people, more than doubling the 2019 figure. This rapid increase is happening at an unprecedented rate and it is expected to escalate even further, especially in developed countries [WHO, (access date: April 16, 2023]. Aging is, of course, not a disease, but it also causes alterations on the speech signal that may either overlap with or mask the effects of certain diseases. For that reason, we included this section on aging and its impact on the speech production process and thus on the speech signal.

2.3.1 The effects of aging on speech

Presbyphonia is the medical term for the changes in the voice due to normal aging [Singh, 2019]. From childhood to adulthood, there are rapid changes in the voice. From the early twenties until about the age of 55, the voice pitch remains relatively stable, after which the vocal tract tissue starts undergoing deterioration, resulting in voice changes. These changes depend from person to person. Singh [2019] describes that some people may develop voice tremors due to neurological changes, some people may experience vocal fold thickening resulting in F0 decrease, while others may experience vocal fold atrophy resulting in F0 increase. The author also describes that in women, pitch tends to decrease with age and in men, pitch tends to increase with age. These changes are associated with the composition of the *lamina propria* of the vocal folds. Some men also experience a thickening of the mucosa, which becomes more viscous and may result in slurred enunciation of speech sounds [Singh, 2019]. The author argues that (quoting) "these changes of voice are continuous, and can often be related to the pitch of the person at or close to the beginning or end of the steady-state for voice, i.e adulthood and middle-age", and thus vocal changes due to aging can be predicted.

Schwoebel et al. [2021] recorded 6650 subjects – most participants reported speaking English as their first language, feeling reasonably well, and being non-depressed and non-anxious – performing different speech tasks, and observed that the gap between F0 of male and female speakers decreases with aging, which is consistent with the F0 variations described by [Singh, 2019]. The authors also observed a decrease of speech rate in free speech and reading tasks, and an increase in pause duration, with aging.

2.3.2 Multimorbidity

The coexistence of two or more chronic conditions in the same individual, or *multimorbidity*, is common and has been rising in prevalence over recent years [World Health Organization, 2016]. A study performed in a developed country found that 25% of the population had two or more long-term conditions [Barnett et al., 2012], and other studies have also shown high levels of multimorbidity in low- and middle-income countries [Wang et al., 2014]. The coexistence of multiple diseases leads not only to increased referrals between healthcare providers, but also to complex drug treatments, which are often based on clinical guidelines defined and tested for single diseases, and thus may result in hazardous outcomes [Ferrucci et al., 2020; World Health Organization, 2016]. The problem of multimorbidity gains special importance in the context of an aging population, where the coexistence of multiple diseases tends to be the norm and not the exception [World Health Organization, 2016].

There are two main mechanisms that mediate the physical decline, the cognitive decline, and frailty, characteristic of aging, which act in opposite directions: one is the accumulation of damage, and the other is the resilience. Accelerated aging may occur either because of faster rates of damage accumu-

lation or because of rapid shrinking and eventual collapse of resilience [Ferrucci et al., 2020]. Resilience is "the harmonic assemblage of the biochemical processes that are aimed at maintaining the identity, integrity, and autonomy of individual organisms against the perturbations induced by both internal and external environments" [Ferrucci et al., 2020].

Some of the most frequent medical conditions in elders include cancer, heart disease, PD, AD, stroke, and arthritis. Furthermore, these medical comorbidities are a risk factor for depression, as well as the converse [Krishnan et al., 2002]. The prevalence of the coexistence of depression and other diseases varies across different studies. Greenwald [1995] reports that the prevalence of major depression and AD is in the range of 5%–15%, and major depression and PD of 15%–20%, with another 25% of patients in each group suffering from minor depression [Krishnan et al., 2002].

Depression and cognitive decline

Depression has been considered a risk factor for the later development of dementia [Krishnan et al., 2002]. In fact, a study including 1764 older adults without baseline cognitive impairment found that depressive symptoms were predictive of cognitive decline [Wilson et al., 2014]. Late-life depression is also associated with cognitive decline, mild cognitive impairment and dementia [Laird et al., 2019]. This association may be due to late-life depression and cognitive decline being manifestations of the same underlying neuropathology. Both conditions are associated with reduced brain volume, increased hippocampal atrophy, increased white matter microstructural changes, and altered structural and functional connectivity [Laird et al., 2019].

The differences in the susceptibility to cognitive decline resulting from aging, pathology, or insult[3] can be explained by the cognitive reserve. Factors such as early-life cognitive ability/intelligence, education, occupation complexity, physical exercise, social and cognitive engagement contribute to cognitive reserve by promoting neuroplasticity. Cognitive reserve also provides psychological resilience [Laird et al., 2019]. Evidence suggests cognitive reserve reduces the association between cognitive impairment and depressed mood, as shown in a study with 37,000 older adults [Opdebeeck et al., 2015].

Depression and Parkinson's disease

A systematic review [Dallé and Mabandla, 2018] focusing on the connections between early life stress, depression, and PD concluded that early life stress may contribute to the development of depression and patients with depression are at risk of developing PD later in life. Depression has been identified both as a risk factor and as a symptom of PD [Dallé and Mabandla, 2018]. Some studies suggest that depression in PD is simply a reaction to the disability of the illness, but research has shown that the incidence of depression in PD is higher than in other illnesses with similar disability [Krishnan et al., 2002].

[3] *Insult* refers to an injury, attack or trauma to the body or one of its parts, or something that causes or has a potential for causing such injury.

Depression is frequently both underdiagnosed or overdiagnosed in PD patients because the typical appearance of PD patients has strong similarities to the appearance of depressed patients. Examples of symptoms common to both diseases are hypomimia (reduced degree of facial expression), hypophonia, psychomotor retardation, loss of energy, loss of appetite, loss of libido, and insomnia.

Sleep disorders, depression and cognitive impairment

There is a bidirectional association between sleep disorders and depression. On one hand, sleep disturbance is the most prominent symptom of depressive patients and it is typically considered a consequence of depression. On the other hand, insomnia was also identified as a risk factor for the development of emerging or recurrent depression among young, middle-aged and older adults [Fang et al., 2019].

OSA has also been associated with affective disorders, and often leads to decline of cognitive functions, or even permanent brain damage. Vanek et al. [2020] conducted a narrative review which analysed 125 papers on the association between OSA, depression and cognitive decline. Although evidence for the link between depression and OSA varies, several of the analysed studies report an interrelationship between the two disorders. The authors describe several possible links between OSA and depression. One possible link is the fact that disturbed sleep patterns negatively impact the stress system, and thus increase the susceptibility of OSA individuals to depression. Another possible link is through obesity: obesity is a major risk factor for OSA and there is an increased prevalence of depression in obese patients. The connection between obesity and depression is bidirectional: obesity can be a risk factor for depression, and weight gain can be a side effect of several psychopharmacs. The third link could be the role of chronic inflammation both in OSA and depressive patients.

OSA's daytime symptoms, such as excessive sleepiness, loss of energy, irritability, withdrawal from social activities, poor concentration, cognitive dysfunction, anxiety or depressive mood problems, and psychomotor changes, are very similar to the symptoms of major depressive disorder. For this reason, assessing the link between depression and OSA can be difficult, and it may lead to misdiagnoses. Unrecognized comorbid OSA in a patient with a psychiatric disorder may lead to inappropriate benzodiazepine medications, which may cause more apneas or hypopneas as they decrease muscle tonus.

The review also found evidence for cognitive impairments caused by OSA, which could be mediated through repetitive hypoxemia. In fact, imaging studies show that part of the damage from untreated OSA can be irreversible. Nevertheless, the study calls for future research, as impairments in cognitive functions could be mimicked by loss of concentration due to sleep deprivation.

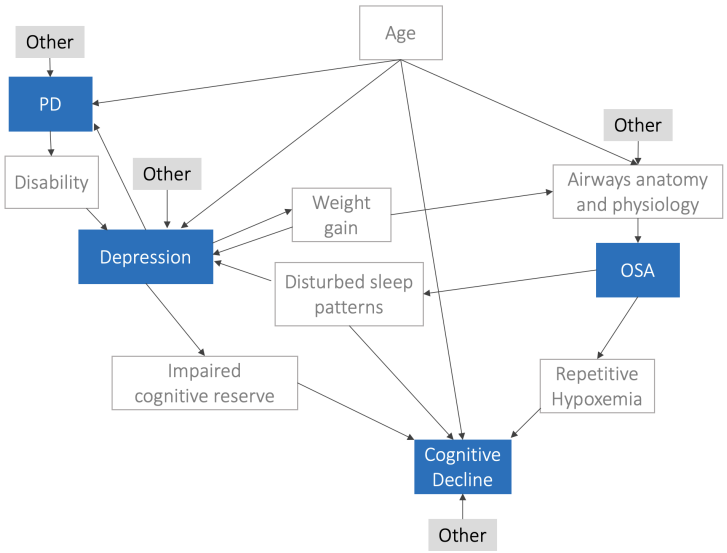


Figure 2.4: Interrelationships between different diseases that affect an aging population.

2.4 The importance of a global perspective on health

The above section describes the interconnections between depression, OSA, AD, and PD. Figure 2.4 presents a diagram that summarizes the primary mechanisms modulating the coexistence of these conditions. The diagram highlights the high likelihood of these diseases co-occurring in the same individuals, as they act as mutual risk factors.

The World Health Organization emphasizes that the overall healthcare of people with multiple health conditions should be provided by medical generalists who combine a community base and comprehensive clinical skills with "interpretive medicine", integrating multiple sources of knowledge with individual needs assessment [World Health Organization, 2016]. Likewise, we hypothesize that a speech-based tool to support medical diagnosis and monitoring of chronic conditions, as is the case of AD, PD, and OSA, should address health from a holistic perspective, and allow an interpretative assessment of multiple diseases, rather than providing a binary classification between a given disease and healthy controls. Moreover, the ongoing diversification of medical disciplines increases the difficulty of identifying all diseases for a single specialist. This further underscores the need for a comprehensive, speech-based diagnostic tool that can assist in the identification and monitoring of multiple conditions.

Besides the fact that these diseases often co-exist, it is also important to note that their manifestations on the speech signal (described in the previous sections) partly overlap with each other. Figure 2.5 presents a diagram where we summarize how each of the described diseases impacts the speech signal, and examples of speech features that can capture such alterations. As can be seen in the diagram, it is notorious that the impact of certain diseases on the speech signal may overlap with that of other diseases. For example, both depression and PD are associated with psychomotor retardation, and thus

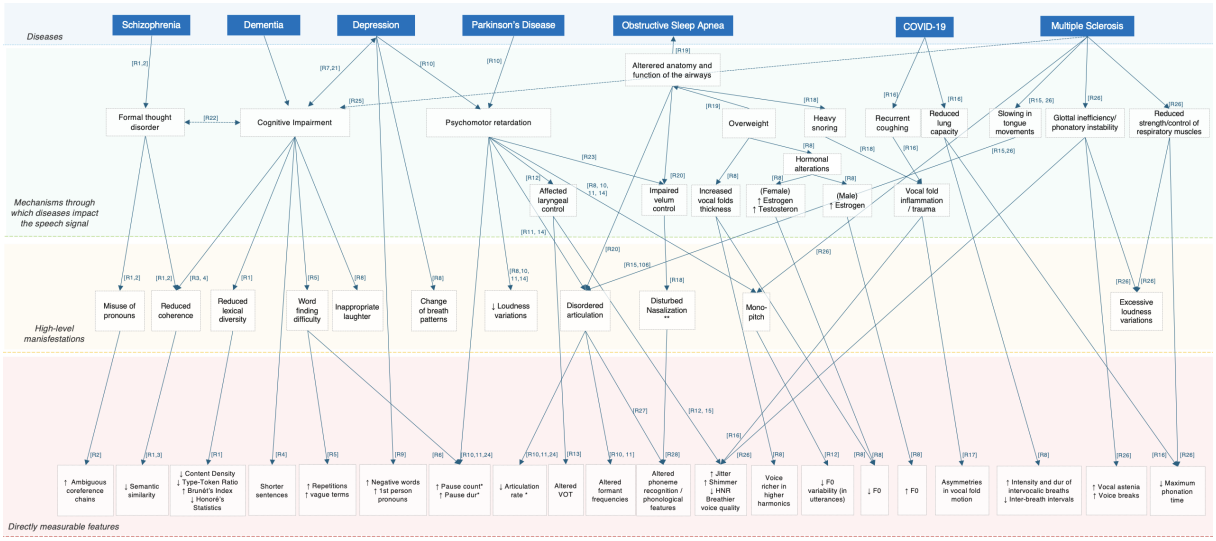


Figure 2.5: Examples of mechanisms through which speech affecting diseases impact the speech signal. The diagram includes references R1–[Voleti et al., 2019], R2–[Bedi et al., 2015; Iter et al., 2018], R3–[Sanz et al., 2022], R4–[Hier et al., 1985; Pompili, 2019], R5–[Forbes et al., 2002; Oppenheim, 1994], R6–[Hoffmann et al., 2010], R7–[Krishnan et al., 2002; Laird et al., 2019], R8–[Singh, 2019], R9–[Rude et al., 2004; Tølbøll, 2019], R10–[Flint et al., 1993], R11–[Cummins et al., 2015a], R12–[Ma et al., 2020], R13–[Vásquez-Correa et al., 2017], R14–[Ramig et al., 2008], R15–[Hecker et al., 2022], R16–[Asiaee et al., 2020], R17–[Al Ismail et al., 2021], R18–[Pozo et al., 2009], R19–[Malhotra and White, 2002], R20–[Monoson and Fox, 1987], R21–[Halahakoon et al., 2019], R22–[Kerns and Berenbaum, 2002], R23–[Hoodin and Gilbert, 1989], R24–[Martínez-Sánchez et al., 2016], R25–[Noffs et al., 2021], R26–[Noffs et al., 2018], R27–[Jiao et al., 2017], R28–[Saxon et al., 2019]. (*) articulation rate and pauses have been reported to be associated with depression, via psychomotor retardation, however, although psychomotor retardation is also present in PD, these features have inconsistent reports for PD [Skodda, 2011]. (**) Disturbed nasalization, as a consequence of impaired velum control, is associated with PD via psychomotor retardation and with OSA. In PD an increase in nasal airflow is reported [Hoodin and Gilbert, 1989], while for OSA, a smaller difference between nasal and oral sounds is reported [Monoson and Fox, 1987].

have similar manifestations on the speech signal. It also becomes clear that often these speech features that capture speech alterations are non-specific for a single disease, and thus, when considered alone, may be insufficient for the automatic detection of diseases. F0-based features, for instance, appear to be altered as a result of several diseases, not to mention possible alterations associated with healthy aging, described earlier in section 2.3.1, or even healthy changes across the menstrual cycle in fertile women [Bryant and Haselton, 2009]. It is also important to refer that some of the features depicted in the diagram have had inconsistent reports in the literature. For example, Voicing Onset Time (VOT) has been found both higher and lower in people suffering from PD when compared with healthy controls [Fischer and Goberman, 2010]. The frequencies of formants have also been inconsistently reported to change with depression [Cummins et al., 2015a]. Furthermore, there are many other factors that directly or indirectly impact the speech production process, and thus introduce alterations on the speech signal. Some of these factors include medication, or other medical interventions, emotions, or mental states.

Considering all these possible sources of variation, it is worth pointing out that the diagram in Fig-

ure 2.5 does not intend to be an exhaustive list of all manifestations of each disease, nor an exhaustive list of all diseases that can impact speech. Instead, with this figure, we aim to provide a rough overview of the most notorious effects in speech associated with each of the diseases explored in this dissertation. In fact, there is a vast literature describing how speech is impacted by individual diseases, but to the best of our knowledge, there are no works that systematize the overlapping effect of multiple diseases in speech together with the main mediating mechanisms.

This diagram provides our effort in this direction, and represents a first step towards a better understanding of how speech could be used as a biomarker for multidisease screening. Singh, in her book [Singh, 2019], defends that causal relationships between parameters and voice must be sought, or reasonably guessed, and then features should be chosen to capture such causal relationships. This diagram intends to allow such reasoning. For example, if we hypothesize that COVID-19 causes vocal fold inflammation due to repetitive coughing, features that capture such inflammation should be derived to predict COVID-19. Furthermore, when studying a different disease, not contemplated in the diagram, one can start by asking if it shares any of the mechanisms already listed, for example, *is the disease often associated with overweight?* or *does the disease cause reduced lung capacity?* In such cases, one can anticipate some of the mentioned speech alterations. It is noteworthy that most of the mechanisms that mediate the diseases' impact on speech signals are hypotheses presented in the literature, and are not necessarily present in all cases of the disease. Further research and discussion within multidisciplinary teams are required to validate these hypotheses.

Another key aspect of this diagram is that it suggests that one possible approach for the use of speech as a biomarker for health conditions is to focus on detecting certain symptoms, risk factors, and/or pathological mechanisms, instead of providing only a binary classification. Unfortunately, to the best of our knowledge, no public datasets exist with such annotations.

2.5 Summary

This chapter provides an overview of the fundamental concepts necessary for understanding the utilization of speech as a biomarker in the detection of various diseases. The discussion encompasses the speech production process, speech-affecting diseases, their interrelationships, and their effects on the speech signal.

This chapter further emphasizes the importance of a holistic perspective to health that transcends simple binary classification of diseases. Indeed, it explores how several speech-affecting diseases and the process of healthy aging act as mutual risk factors, sometimes influencing speech in overlapping ways.

3

Background: automatic detection of speech affecting diseases

Contents

THIS chapter presents the corpora and the methods typically used for the automatic detection of speech affecting diseases. In particular, it introduces frequently used features, strategies for working with limited-size corpora, and evaluation metrics. Further detailed literature review is provided throughout the dissertation, as relevant to each chapter.

3.1 Pipeline for the automatic detection of speech affecting diseases

The typical approach for disease detection from speech signals, summarized in Figure 3.1, follows the general pipeline used to solve classification problems using machine learning. It starts with a data collection step, which is often conducted in partnership with healthcare institutions that can provide and annotate the data. Afterwards, there is a pre-processing step that can include silence removal, diarization, audio segmentation, and transcription of the acoustic signal into text records. Although datasets are frequently noisy, it is not advised to perform algorithmic data enhancement, as most noise removal techniques modify the signal and leave artifacts in it, which are likely to modify the micro-traces in the signal that are necessary for disease detection [Singh, 2019]. After data pre-processing, data are represented in a format that can be used as input to the machine learning systems: either via knowledge-based features[1] or representations learnt by a neural network. In the context of disease detection from speech, datasets are often small and the feature sets have high dimensionality. Thus, there may be a step of dimensionality reduction to avoid the *curse of dimensionality* and overfitting. Finally, there is a classification stage. Most studies are designed to solve a binary classification task, aiming at distinguishing a given disease from healthy controls. Besides classification, several works also focus on the regression of disease severity, which is particularly importance for monitoring disease progression. The scenario of multi-disease detection is rarely addressed.

Some of the data representation approaches leverage the concept of transfer learning, capitalizing on the enormous potential of large deep learning models at the feature extraction stage. Conversely, the classifiers used for disease detection are frequently traditional machine learning algorithms, such as Support Vector Machine (SVM), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Gaussian Mixture Model (GMM), Probabilistic Linear Discriminant Analysis (PLDA), Decision Tree (DT), or Random Forest (RF) which can be more effectively trained using smaller amounts of data. The state-of-the-art using deep learning for speech based classification tasks, and particularly for health, is evolving at an incredible pace. However, these methods demand extensive data, which is often unavailable for speech-based disease detection.

[1] *knowledge-based features* refer to features chosen and derived based on domain-specific knowledge, e.g. by experts who have a deep understanding of the specific domain, or based on the literature.

Figure 3.1: General pipeline for the automatic detection of diseases, from speech.

3.2 Speech representation

There is a plethora of speech data representations described in the literature, used for distinct applications. In this section, we describe several of these techniques, focusing on the techniques used throughout the experiments described in this thesis. Providing a comprehensive listing and description of all speech representation techniques would be impractical due to the vast scope and extensive nature of the subject.

MFCC Mel-frequency cepstral coefficients (MFCC) represent the envelope of the short time power spectrum, which is a manifestation of the vocal tract shape. The Mel scale simulates the human auditory canal, as the frequency bands become wider for higher frequencies. MFCC are a standard representation of the speech signal, used in many applications, including as input features to neural networks trained for the extraction of neural representations, as is the case of x-vectors, described below.

ComParE The Computational Paralinguistics Evaluation (ComParE) feature set [Eyben et al., 2013; Schuller et al., 2013b] consists of 6373 features, originally proposed for the Computational Paralinguistics Challenge 2013. It was designed based on the features used for the challenges 2009-2012, and the lessons learnt in those challenges. It includes energy, spectral, MFCC, and voicing-related low-level descriptors. These low-level descriptors include logarithmic HNR, voice quality features, Viterbi smoothing for F0, spectral harmonicity, and psychoacoustic spectral sharpness. Statistical functionals are also computed. This feature set has been widely used in the speech paralinguistic community throughout the years, and it continues to be used recently (e.g. for AD detection [Luz et al., 2020]). Its implementation is publicly available with the openSMILE toolkit [Eyben et al., 2013].

eGeMAPS The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), proposed in 2016, was designed to be a basic standard acoustic parameter set for various areas of automatic voice analysis, such as paralinguistic and clinical speech analysis, based on their theoretical significance, their capability to reveal physiological alterations, and the good results they obtained in previous works [Eyben et al., 2016]. eGeMAPS is a feature set of 88 low-level descriptors and functionals that represent the speech signal in terms of spectral, cepstral, prosodic, and voice quality parameters. Its implementation is publicly available with the openSMILE toolkit [Eyben et al., 2013].

i-vectors The i-vector approach [Dehak et al., 2010] was first introduced as a speaker representation for speaker verification. It aims at modeling together speaker and channel variability in a single low-rank

sub-space, called the total variability space. It has been successfully used to solve other tasks related to speech, including language recognition [Dehak et al., 2011] and emotion recognition [Lopez-Otero et al., 2014]. More recently, i-vectors have also been used for disease detection from speech (e.g. for PD detection [Moro-Velazquez et al., 2020], and OSA detection [Perero-Codosero et al., 2019]).

x-vectors x-vectors are deep neural network based speaker embeddings, proposed as an alternative to i-vectors for speaker [Snyder et al., 2018] and language recognition [Snyder et al., 2017b] tasks. The neural network used to produce the x-vector embeddings comprises three main blocks. The first block is a set of time delay neural network (TDNN) layers operating at frame level with a small temporal context. These layers work as a 1-dimensional convolution with a kernel size corresponding to the temporal context. The second block, a statistical pooling layer, aggregates the information across the time dimension and outputs a summary for the entire speech segment that captures long-term speaker characteristics. The third block is a set of fully connected layers, from which x-vectors embeddings can be extracted after the network is trained. Contrarily to i-vectors, which are generative, the x-vector system is discriminative, hence questions could be raised on whether it would also capture paralinguistic information. It has been shown that it does carry paralinguistic information, in particular the health state of the speaker. X-vectors have been used for the automatic detection of OSA [Perero-Codosero et al., 2019], PD [Correia et al., 2021; Moro-Velazquez et al., 2020], AD [Zargarbashi and Babaali, 2019], and depression [Correia et al., 2021]. These embeddings can be extracted using pre-trained models or trained from scratch using the Kaldi toolkit [Povey et al., 2011], or more recently SpeechBrain [Ravanelli et al., 2021]. Since x-vectors first appeared, several enhancements have been proposed to the original architecture. In particular, we highlight the factorized time delay neural network (TDNN-F) architecture proposed by [Povey et al., 2018] for speech recognition, and by [Villalba et al., 2020] for speaker recognition; and the Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Networks (ECAPA-TDNN) architecture, introduced by [Desplanques et al., 2020], with enhancements based on the recent trends in the computer vision field.

PASE+ The problem-agnostic speech encoder (PASE) model, introduced by [Pascual et al., 2019] and enhanced to PASE+ by [Ravanelli et al., 2020], is a self-supervised encoder for robust representation learning. PASE+ combines a convolutional encoder followed by twelve workers, that cooperatively solve various self-supervised regression and binary classification tasks. The regression tasks include the reconstruction of the waveform, log power spectrum, MFCCs, prosody, filter banks, and gammatone. This reconstruction is made in windows of 25 ms and 200 ms. The binary workers are trained to maximize mutual information between anchor and positive samples from the pool of PASE-encoded representations, thus aiming at extracting higher-level information from speech signals. Two primary tasks are defined for the binary workers, based on different sampling strategies: *Local Info Max*, which samples the positive representation from the same sentence as the anchor and the negative from a different sen-

tence, likely from another speaker, encouraging the model to distinguish speakers based on consistent local features; and *Global Info Max* for which the anchor and positive representations are obtained by averaging all the PASE features extracted over long chunks of 2 s belonging to the same sentence, while the negative representation is obtained over a chunk of 2 s from a different sentence.

PASE+ features are extracted at the end of the encoder after joint training of the encoder and the workers. These features have been shown to contain relevant information for paralinguistic tasks such as emotion recognition [Pascual et al., 2019].

Many other recent speech representations have been used for disease detection that are not thoroughly described here as they were not employed in the experiments conducted in this work, however, they deserve mention. These include the Bag-of-Audio-Words [Riley et al., 2008; Schmitt et al., 2016], Deep spectrum [Amiriparian et al., 2017], auDeep [Freitag et al., 2017], TRILL (TRIpLet Loss network) [Shor et al., 2021], Bidirectional Encoder Representations from Transformers (BERT) embeddings [Devlin et al., 2018], wav2vec 2.0 embeddings [Baevski et al., 2020], HuBERT embeddings [Hsu et al., 2021a], and WavLM embeddings [Chen et al., 2022]. Wav2vec 2.0 and WavLM representations are used in this thesis's experiments to perform automatic speech recognition, rather than directly as features for disease detection. Unlike the other representations mentioned here, BERT embeddings are derived from text, but they have been extensively used on top of manual or automatic transcriptions for disease detection (e.g. [Pompili et al., 2020b]). It is worth noting that most of these models leverage self-supervised learning, which allows them to learn rich representations from vast amounts of unlabeled speech data, making them particularly powerful for various downstream tasks.

3.3 Existing Corpora

This section presents an overview of the speech corpora designed to study speech affecting diseases, which have been used throughout the experiments described in this thesis. In particular, we will cover the in-the-Wild Obstructive Sleep Apnea Corpus (WOSA), the Parkinson's disease Corpus from the Applied Telecommunications Group (GITA) at the Universidad de Antioquia, Colombia Parkinson's Spanish corpus (PC-GITA), the in-the-Wild Speech Medical Corpus (WSM), the Distress Analysis Interview Corpus – Wizard Of OZ (DAIC-WOZ), the Alzheimer's Dementia Recognition through Spontaneous Speech (corpus) (ADReSS), the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE), the COVID-19 COUGH corpus (C19C), the COUGHVID crowdsourcing dataset, and the Crowdsourced Language Assessment Corpus (CLAC). Table 3.1 summarises the main characteristics of these corpora. Additionally, we describe VoxCeleb and the Texas Instruments/Massachusetts Institute of Technology corpus (TIMIT), which, although not originally designed for disease detection, were used in some of the experiments conducted in this thesis.

Table 3.1: Corpora of speech affecting diseases. *m*, *f*, and *o* refer to genders – male, female and other. *Although WOSA corpus was not made publicly available, the vlogs can be found on YouTube. †ILSE is a longitudinal corpus, thus the number of participants with disease corresponds to the number of participants that have any type of cognitive impairment (AD, age-associated cognitive decline, or other forms of dementia) at any of the recording times. ‡ information not made available.

Corpus	Disease	Language	Publicly Available	Tasks	Participants (m; f; o)		Observations
					Control	Disease	
PC-GITA	PD	Sp	Yes	exercises	25; 25; 0	25; 25	–
WSM_{PD}	PD	Eng	Soon	vlog	98; 106; 0	105;104	–
WSM_{depr}	depression	Eng	Soon	vlog	130; 146; 0	123; 144	–
DAIC-WOZ	depression	Eng	Yes	interview	61; 39; 0	18;24	–
WOSA	OSA	Eng	No*	vlog	4; 4; 0	4; 4; 0	–
ADReSS	AD	Eng	Yes	image description	35; 43; 0	35;43	–
ILSE	AD	Ger	No	interviews	349; 338; 0	171; 144†	longitudinal
C19C	COVID-19	-	Yes	cough	‡	‡	cough dataset
COUGHVID	COVID-19	-	Yes	cough	5632; 2638; 0	395; 285	cough dataset
CLAC	healthy	Eng	Yes	exercises	903; 916; 13	–	healthy only

Parkinson's disease Corpus from the Applied Telecommunications Group (PC-GITA)

The PC-GITA [Orozco-Arroyave et al., 2014], also referred to as the New Spanish Speech Corpus in some publications, is a collection of speech recordings, where the subjects perform a number of speech exercises. This dataset is balanced both in terms of gender and age, and includes subjects with ages ranging from 31 to 86 years old. The corpus is spoken in Colombian Spanish, and the recordings were captured in noise-controlled conditions, in a sound-proof booth that was built at the Clínica Noel, in Medellín, Colombia. Participants were diagnosed by neurology experts, and were labeled according to standard clinical protocols: the UPDRS, and Hoehn and Yahr (H&Y) [Goetz et al., 2004]. The recording protocol considered different tasks, including repeating and sustaining vowels; dysdiadochokinesia analysis, or DDK (i.e. the rapid repetition of words and phonemes such as */pa-ta-ka/*, */pa-ka-ta/*, */pe-ta-ka/*, */pa/*, */ta/*, */ka/*); repeating sentences with different levels of syntactic complexity; reading sentences and dialogues, and spontaneous speech. The complete evaluation protocol amounts to less than 10 minutes of speech per patient. Each of the subjects completed the full battery of exercises, for a total of 4800 recordings. The corpus was made available for the 2015 ComParE challenge [Schuller et al., 2015], for the regression of PD severity.

In-the-Wild Speech Medical corpus (WSM)

The WSM corpus [Correia, 2021; Correia et al., 2021] comprises 928 vlogs from YouTube, totaling over 131 hours of speech from individuals potentially affected by depression (WSM_{depr}) or Parkinson's disease (WSM_{PD}). WSM is thus an audiovisual corpus, containing recordings of subjects speaking spontaneously about topics of their choosing, without any guided exercises or interviews, and under varying channel and noise conditions, contrasting with typical speech-affecting disease corpora.

Videos were annotated based on the presence of a self-reported health status claim, verified by crowdsourced manual inspection. Thus, labels are not medically validated and may be noisy. Never-

theless, evidence suggests that self-reports are reliable proxies for actual health status [Correia, 2021; Correia et al., 2021]. Each video is associated with crowdsourced annotations not only for the self-reported health status in terms of the target disease, but also for the perceived age and gender of the speakers. WSM corpus is balanced in terms of perceived age, gender and self-reported diagnosis.

DAIC-WOZ

The DAIC-WOZ [Gratch et al., 2014] is a multi-modal collection of semi-structured clinical interviews performed by an animated virtual agent, controlled by a human interviewer out of the participants sight. It was designed to simulate the standard protocols created for identifying people at risk for depression, anxiety and post-traumatic stress disorder. The participants were labelled for depression using the Patient Health Questionnaire (PHQ-8) [Kroenke and Spitzer, 2002] . The age distribution is not reported. This corpus was made publicly accessible, through the Audio/Visual Emotion Challenge and Workshop (AVEC 2016) [Valstar et al., 2016], and its subsequent editions. The DAIC-WOZ contains 189 interviews, ranging from 5 to 20 minutes, which were segmented into utterances at boundaries with at least 300 ms of silence. Out of the 189 interviews, 45 interviews were made available as test set, without label annotations. Thus, the number of participants in Table 3.1 excludes these 45 interviews.

In-the-Wild Obstructive Sleep Apnea (WOSA) Corpus

The WOSA corpus is a pilot corpus introduced in our earlier work [Botelho, 2018; Botelho et al., 2019], inspired by the concept of the WSM corpus [Correia et al., 2018a]. This corpus consists of segments from 16 YouTube vlogs, all featuring English-speaking individuals. The dataset includes eight subjects who claim to suffer from OSA and eight control subjects. Among the OSA subjects, six were undergoing treatment with continuous positive airway pressure (continuous positive airway pressure (CPAP)) during sleep, one was using an oral appliance, and one was not receiving any treatment. The control subjects were selected randomly from vlogs on unrelated topics. Later in this thesis (chapter 5), we expand the WOSA corpus to encompass 40 subjects.

ADReSS

The ADReSS corpus [Luz et al., 2020] comprises speech recordings and the corresponding manual transcriptions of 156 subjects describing the Cookie Theft picture. This dataset is a subset of the Pitt corpus [Becker et al., 1994], also known as DementiaBank, and was curated for the 2020 ADReSS Challenge [Luz et al., 2020]. The participants include 78 individuals diagnosed with Alzheimer's disease and 78 healthy controls, matched for age and gender. Speech recordings were segmented using voice activity detection and subsequently normalised. The dataset made available contained both full enhanced audio, and normalised audio chunks. The primary objective of this dataset is to address the lack of standardisation that currently affects the field of dementia detection from speech, facilitating systematic

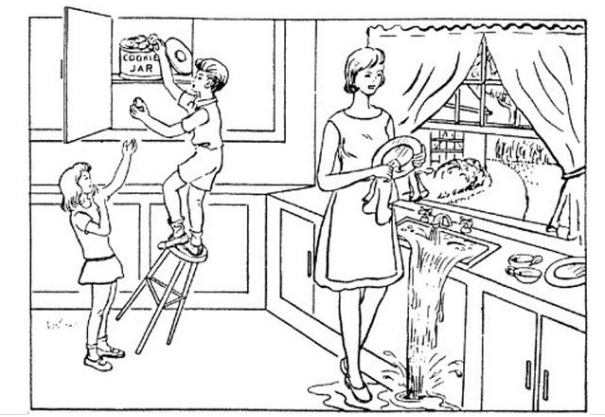


Figure 3.2: The Cookie Theft picture, from the Boston Diagnostic Aphasia Examination [Goodglass et al., 2001].

comparison of various approaches.

The "Cookie Theft" image, depicted in Figure 3.2, has been widely used for studying cognitive impairments through speech. It can be described using seven concepts: woman doing dishes, sink overflowing, boy on stool, children stealing cookies, girl reaching for cookie, stool falling, and woman not noticing [Mackenzie et al., 2007; Nicholas and Brookshire, 1995].

ILSE

ILSE [Sattler et al., 2015] is a longitudinal corpus, designed to investigate satisfying and healthy aging in middle adulthood and later life. It includes participants born in Germany, in two distinct cohorts: 1930-1932 and 1950-1952. Each participant engaged in up to four measurements, conducted in intervals of approximately 5 years. At each measurement, the subjects were recorded during biographic interviews, where they gave elaborate answers to open-ended questions. During the interviews, the participants had enough time to think about their answers. The average duration of the recordings became shorter with each follow-up interview, given that with each measurement time, the gathered biographic information was accumulated and the questions were reduced. The speech recordings of the first two measurements were stored on tapes. For the third and fourth measurement, digital recording devices were used. All interviews have been digitized using a sampling rate of 16kHz, and 16-bit linear PCM quantization. The changing recording conditions over time are one of the main challenges of this corpus. Further details about the ILSE corpus can be found in [Martin et al., 2000].

The effort to manually transcribe the interviews is ongoing. Currently, there are manual transcripts for 145 interviews from 91 participants. These transcripts have varying quality and do not include time alignment information. To distinguish this subset of the ILSE corpus from the subsets used in previous works, and from the subsets that may be used in the future with larger the dimensions, we denote this subset as $\text{ILSE}_{\text{m145}}$. Out of these 145 interviews, 108 are controls, 16 suffer from AD, and 21 have age associated cognitive decline.

The entire ILSE corpus is fairly large, when compared to other health-related speech corpus, as it consists of approximately 6,500 hours of recordings, including over 1000 participants, recorded over a 20 year period. However, the number of participants with AD is limited because it represents the natural prevalence of dementia in the population studied. In fact, only 27 out of the 1002 develop AD. Due to the coverage of sensitive private information, the corpus is not publicly available.

COVID-19 Cough Corpus

The C19C corpus is a curated subset of the Cambridge COVID-19 Sound database [Brown et al., 2020; Han et al., 2021], a crowdsourced corpus with examples of breathing, coughs and speech recorded in-the-wild, made available for the Computational Paralinguistic Challenge in 2021 [Schuller et al., 2021]. The C19C corpus contains 725 cough recordings from 397 participants and the corresponding labels for the self-reported COVID-19 status. Data are distributed in three speaker-independent, gender-balanced subsets: train (71 positives/215 negatives), development (48 positives/183 negatives) and a blind test set (208 samples).

COUGHVID

The COUGHVID corpus [Orlandic et al., 2020] is a non-curated, publicly available dataset[2]. Recordings were performed using a lossy codification and present a variety of conditions (sampling rate, bandwidth, number of channels, and quality). Volunteers recorded their coughs and reported their COVID-19 status (positive/symptomatic/healthy), age, gender, and medical condition. A small fraction of the dataset was annotated by expert pulmonologists with information regarding the type of cough (wet/dry/inconclusive), presence of audible dyspnea, wheezing, stridor, choking, and nasal congestion, diagnosis (upper/lower respiratory tract infection/obstructive lung disease/COVID-19/healthy cough), and severity (healthy cough/mild/severe). It comprises 27,550 recordings, 15,125 of which are classified as coughs by an automatic cough detector developed by the COUGHVID team. 10,763 of them have self-provided gender and COVID-19 status annotations, distributed as follows: 680 COVID-19 positives (395 male/285 female), 8270 healthy (5,632 male/2638 female), and 1,813 symptomatic (1,114 male/699 female). Table 3.1 only includes the COVID-19 positive and healthy people.

CLAC

The Crowdsourced Language Assessment Corpus (CLAC) [Haulcy and Glass, 2021] was created to provide a collection of audio samples from various healthy speakers, providing speech tasks similar to what is standard in corpora for the detection of speech affecting diseases. The authors suggest that this corpus can be used to learn general representation of speech from healthy subjects. CLAC includes speech from 1,832 speakers almost all located in the USA. The speakers were recruited via

[2] https://zenodo.org/record/4498364

a crowdsourcing platform and claim to have no health-related symptoms that might affect their speech. The speech tasks include counting from 1 to 20, saying the days of the week, describing two images – Cookie theft, and Picnic – reading two passages, repeating the words “artillery”, “catastrophe” and “impossibility”, and maximum phonation of the vowel /a/. In addition to the audio files, the corpus includes transcripts for several tasks, as well as speaker metadata, including age and years of education. The latter is a pertinent variable, especially when investigating disorders that affect discourse. The average age of the participants was 35.7 years and the average years of education was 15.4. Transcripts were automatically generated using the Google Speech Recognition API [Zhang, 2017].

TIMIT

The TIMIT corpus [Garofolo et al., 1992] contains read speech recordings from 630 native speakers of American English representing 8 major dialect divisions, with each speaker reading ten phonetically-rich sentences. The speakers were screened by a professional speech pathologist. One subject was excluded due to lack of age information. The corpus was designed for automatic speech recognition.

VoxCeleb

VoxCeleb [Nagrani et al., 2019] is a large-scale corpus intended for speaker verification, containing short clips from celebrity interviews uploaded to YouTube. The corpus is composed of two parts, VoxCeleb 1 and 2, totalling over 7,000 speakers of multiple ethnicities, accents, occupations and age groups.

A subset of VoxCeleb was annotated by Hechmi et al. [2021] with age, gender, and nationality, among other information. This subset includes 840 speakers from the USA, with available age information. Speakers form other countries were not considered to avoid including interviews not spoken in English.

3.4 Coping with small datasets

As described in the previous section, corpora designed for the detection of speech affecting diseases are typically small when compared to corpora used for other speech-based applications, such as speech recognition or speaker recognition, namely for English. Although it would be desirable to work with larger datasets, there are some strategies that have been used to cope with small datasets, in the context of speech affecting diseases.

Data augmentation There are two main approaches to perform data augmentation. The first is to introduce copies of the data with perturbation, e.g. introducing background noise, such as music, babble, or reverberation; speeding up or slowing down the recordings; specAugment [Park et al., 2019]; FilterAugment [Nam et al., 2022]; and FrAUG [Ravi et al., 2022]. This type of strategy is mostly used associated with deep neural networks, and besides increasing the amount of data, it also makes the models more robust to slight shifts in the training domain. The second approach is to synthesize new data using

Table 3.2: Definition of TP, TN, FP and FN.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Generative Adversarial Networks (GANs), as done by [Deng et al., 2017] for autism detection.

Transfer learning This technique consists of training a neural network on a task for which there is a sufficient amount of data, and transfer the learnt representations to a similar task, for which less data is available. A comprehensive survey of various transfer learning approaches is provided in [Zhuang et al., 2020], which categorizes the techniques formally. A large portion of very recent works for disease detection from speech leverages the idea of transfer learning. For instance, studies employing the neural data representations described in the previous section (3.2) for disease detection exemplify this approach. Pompili et al. [2020b], for example, leverages *x-vectors* and *BERT embeddings* for Alzheimer's disease detection.

Intelligent labeling paradigms This technique refers to leveraging a small set of labeled data to annotate a larger dataset with minimal human involvement. Examples of such paradigms include semi-supervised learning, active learning, and cooperative learning [Cummins et al., 2018].

Crowdsourcing Another possibility is to use crowdsourcing to collect new data or to annotate it, e.g. [Brown et al., 2020; Correia et al., 2021; Han et al., 2021; Haulcy and Glass, 2021; Sharma et al., 2020]. The advantage of this approach is that it is able to retrieve new and real data. The main disadvantage in the context of speech affecting diseases is that labels are not medically verified. Besides, the recording conditions typically present large variability.

Cross-validation Cross-validation is a technique for training and evaluating machine learning systems when little amount of data is available [Refaeilzadeh et al., 2009]. It is very frequently used in the context of the detection of speech affecting diseases for that exact reason. However, it may also lead to overfitting to the testing folds and the underestimation of the errors [Vabalas et al., 2019; Varoquaux, 2018].

3.5 Model evaluation and metrics

Several metrics for evaluating classification problems rely on the concepts of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), as defined in Table 3.2.

In the context of binary classification of speech affecting diseases, the most used metric is Unweighted

Average Recall (UAR). This metric, defined in Equation 3.1, assigns equal weight to all classes, making it particularly suitable for evaluating problems with imbalanced class ratios, as is often the case with datasets related to speech-affecting diseases [Cummins et al., 2018]. In earlier paralinguistic studies, the term "unweighted accuracy" was used to refer to UAR (e.g. Schuller et al. [2011]).

UAR, is the average of true positive rates (TPR), also known as *recall* or *sensitivity*, and true negative rate (TNR), also called *specificity*, both of which are more frequently reported in medical literature. Sensitivity (Equation 3.2) describes the percentage of individuals with the disease that are correctly identified by the test, while specificity (Equation 3.3) refers to the percentage of individuals without the disease correctly labelled as controls. Let us consider a scenario involving large-scale screening for a hypothetical highly contagious disease. At the end of the screening process, if the test is negative, the individual is told to continue regular daily activities, whereas a positive result prompts further medical testing due to the risk of having the disease. In this scenario, it is crucial to minimize false negatives, as failing to identify a contagious individual could lead to further spread of the disease. Hence, a very high recall or sensitivity is desired. Conversely, a high number of false positives would overload healthcare services with unnecessary additional tests, consuming valuable resources. This scenario illustrates the trade-off between different metrics and highlights the importance of considering multiple metrics simultaneously, when possible.

Other frequently reported metrics include accuracy, precision, and F1-score, defined in Equations 3.4, 3.5, and 3.6, respectively. It is also common to report unweighted averages of precision and F1-score for both classes.

$$UAR = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{FP+TN}), \tag{3.1}$$

$$TPR = \frac{TP}{TP+FN}, \tag{3.2}$$

$$TNR = \frac{TN}{FP+TN}, \tag{3.3}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}, \tag{3.4}$$

$$Precision = \frac{TP}{TP+FP}, \tag{3.5}$$

$$F1score = 2\frac{precision \times recall}{precision + recall}, \tag{3.6}$$

Besides evaluating model performance, it is also important to evaluate its *robustness*, especially in high stake applications, such as the screening of diseases. Although this type of evaluation techniques is rarely employed in works that aim to perform disease detection from speech, we believe they also deserve mentioning. The *robustness* of the model can be evaluated with the same predictive performance metrics, e.g. UAR but using out-of domain distributions. This can be achieved using, for example, a different corpus for the detection of the same speech affecting disease. However, these corpora are scarce, and even when more than one corpus is publicly available for the same speech affecting disease, they often differ in terms of recording conditions, languages, demographics and speech tasks. This makes measuring robustness more challenging, as it will be discussed in chapters 6 and 7.

3.6 Summary

This chapter provides an overview of the standard techniques used for the automatic detection of speech affecting diseases, as well as the corpora available to the research community. It emphasizes that one substantial challenge in using speech as a biomarker for disease detection is data scarcity. We present strategies to cope with limited-size datasets.

4

Towards silent paralinguistics: EMG

Contents

EMG signals recorded during speech production encode information on articulatory muscle activity and also on the facial expression of emotion, thus representing a speech-related biosignal with strong potential for paralinguistic applications. In this chapter, we estimate the electrical activity of the muscles responsible for speech articulation directly from the speech signal. To this end, we first perform a neural conversion of speech features into electromyographic time domain features, and then attempt to retrieve the original EMG signal from the time domain features. The results reported in this chapter, which were published at Interspeech 2020 [Botelho et al., 2020a], together with the results present in [Diener et al., 2020] laid the groundwork for the emerging field of Silent Computational Paralinguistics.

This chapter reflects the initial direction of this PhD, which aimed at understanding the interconnections of speech production, the different biosignals that precede the sound wave, and the mechanisms through which each disease impacts speech production. However the generalized COVID-19 lockdown precluded further data collection, hindering this research line. The following chapters shift the focus to other biomarkers that can be collected non-invasively, unobtrusively, and remotely via web or mobile phone applications.

It is important to note that the state-of-the-art mentioned in this chapter reflects the time of publishing these results. The field of silent paralinguistics and silent communications has since seen significant advancements and evolution.

4.1 Introduction

As previously described in chapter 2, not only the articulatory system, but also the respiratory system and the nervous system play an important role in the speech production process. At the different levels of speech production, e. g. at the brain, at the peripheral nervous system, muscular action potentials, or directly during speech kinematics, biosignals can be captured and studied to draw conclusions about linguistic and paralinguistic content of spoken communication [Schultz et al., 2017]. Many researchers have taken advantage of biosignals, proposing systems to generate speech features from Electrocorticography (ECoG) [Angrick et al., 2019; Anumanchipalli et al., 2019; Herff et al., 2016], EEG [Krishna et al., 2019, 2020b], EMG [Diener et al., 2018; Janke and Diener, 2017; Wand et al., 2018], ultrasound [Denby and Stone, 2004; Kimura et al.], and video recordings of speech articulation [Michelsanti et al., 2020; Vougioukas et al., 2019].

The inverse problem of transforming acoustic speech signals into the underlying biosignals involved in speech production has likewise sparked recent interest. In particular, this concerns the issue of acoustic-to-articulatory inversion (AAI) , i. e., the estimation of articulatory movements from the acoustic speech signal. Most AAI works are based on Electromagnetic Articulography (EMA) data (e. g. [Illa and Ghosh, 2019, 2018; Liu et al., 2015]), and ultrasound imaging [Porras et al., 2019]. Krishna et al. [2020a]

converted acoustic features to EEG features, whereas Anumanchipalli et al. [2019] estimates articulatory dynamics from audio recordings, which are later included as an explicit intermediate representation in the decoding of speech from signals captured with ECoG.

By examining how speech is converted into other biosignals involved in speech production, work on AAI may help us understand the interconnections between different biosignals and their role in communication. Among these, signals captured using facial EMG are arguably one of the most important involved in speech production, e. g., due to their role for signalling facial expressions of emotion [Kappas et al.] as well as social intentions such as politeness [Küster, 2020]. Furthermore, the synthesis of EMG from acoustic speech shows rich potential – for example in medicine, as a means to increase articulatory awareness in speech therapy, or in computer animation, as a means to visualize realistic muscle movements [Sagar and Scott, 2009]. AAI research further complements work on Silent Computational Paralinguistics, i. e., the assessment of speaker states and traits from non-audibly spoken communication [Diener et al., 2020], by generating large amounts of synthetic EMG data from audio. Thus, future work on EMG-based speech models may require smaller amounts of costly laboratory recordings once important features can be validated against synthetic EMG obtained from AAI.

To the best of our knowledge, our work is the first that proposes the conversion of acoustic speech into signals captured using EMG. Ours is a two-step approach (see Figure 4.1), motivated by the standard two-step speech synthesis methodologies. First, we generate EMG time domain (TD) features, and then we derive the EMG from those features. In this initial study, we consider these two steps as independent tasks, and consequently, when synthesizing the EMG signal from the TD EMG features, the TD EMG features considered are the ones obtained from the true EMG signal rather than the ones synthesized in step one. Thus, this second step is designed as a proof-of-concept to validate whether the EMG TD features encode sufficient information to retrieve the original EMG signal. With these two separate steps, we intend to present the basis for a single pipeline that allows the retrieval of the original EMG from speech, and also to validate the use of TD EMG features as intermediary representations in future silent paralinguistics systems. For step one, we propose an hourglass-shaped feed forward neural network, while for the second step, we propose a convolutional block followed by a bidirectional long short-term memory (BLSTM) block. Both steps are evaluated using the Concordance Correlation Coefficient (CCC) [Lin, 1989].

4.2 Corpus

All experiments in this study were conducted on the EMG-UKA parallel EMG-Speech corpus [Wand et al., 2014], available from ELRA [ELRA Catalogue ID ELRA-S0390, 2014]. The corpus includes 63 small and large sessions from 8 speakers, in 3 speaking modes (audible, silent speech, and whispered

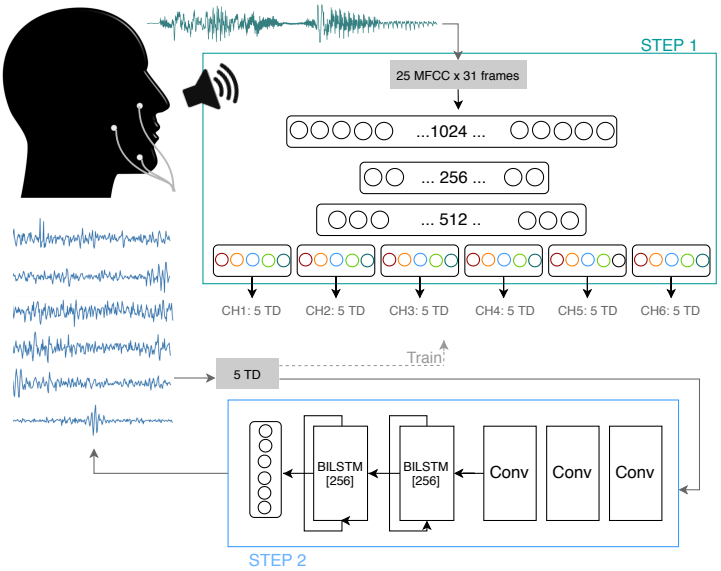


Figure 4.1: Two-step Speech-to-EMG system: $\text{Acoustic}_{\text{MFCC}}$-to-$\text{EMG}_{\text{TD}}$ (step 1) followed by EMG_{TD}-to-EMG_{Orig} (step 2).

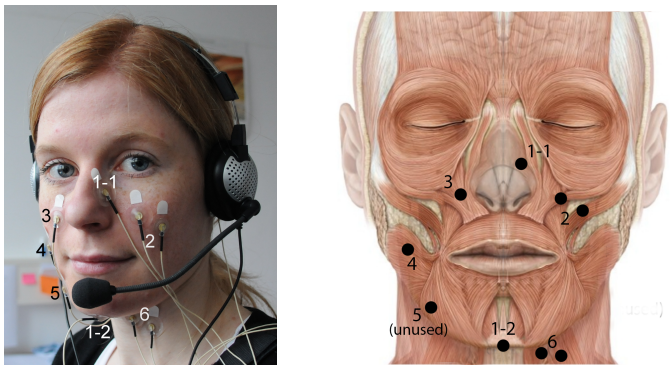
speech). In this work, we use the EMG and speech recordings that correspond to the audible speech mode. Information on the number and duration of the sessions can be found in Table 4.1. Each speaker has a varying number of sessions: two speakers (speaker 2 and 8) recorded a larger number of sessions (32 and 19, respectively) while the other speakers recorded up to 3 sessions. Further information on the sessions can be found in [Wand et al., 2014].

Table 4.1: EMG-UKA Corpus.

Session Type	Number of sessions	Duration [h:m:s]	
		Average	Total
Small	61	0:03:08	3:11:34
Large	2	0:27:02	0:54:04

The acoustic data in the EMG-UKA corpus was recorded at a sampling rate of 16 kHz with a standard close-talking microphone, whereas the speech-related EMG signals were recorded using a Becker Meditec Varioport amplifier with 6 EMG channels, operating at 600 Hz. The two signals were synchronized via a hardware marker that marks the same point in time, and assuming an electromechanical delay between muscle activation and speech production of 50 ms. Figure 4.2 shows the positioning of the electrodes, capturing the EMG signal of six articulatory muscles [Wand et al., 2014]: Zygomaticus major and levator anguli oris (both 2, 3), platysma (4, 5), depressor anguli oris (5), the anterior belly of the digastric (1-2), the tongue (1-2, 6), and a reference channel on the nose (1-1).

Each session is divided into train and test data. The small sessions contain 40 train utterances and 10 test utterances. The large session of speaker 2 contains 500 train utterances and 20 test utterances, and the large session of speaker 8 contains 496 train utterances and 13 test utterances. While the training



(a) (b)

Figure 4.2: EMG electrode positioning in the EMG-UKA corpus. Electrodes numbered in black in (a) are measured against a reference electrode behind the ear, whereas white numbers indicate bipolar derivation. Figure (a) from [Diener, 2021] and muscle chart in (b) from [Wand et al., 2014] and adapted from [Schünke et al., 2006].

data partially varies across sessions, the 10 test utterances are unique and the same in all sessions. In the larger sessions, the test set includes repetitions of the 10 test utterances [ELRA Catalogue ID ELRA-S0390, 2014].

4.3 Method

We address the conversion of acoustic speech to EMG as a two-step problem. In the first step, named "$\text{Acoustic}_{\text{MFCC}} \rightarrow \text{EMG}_{\text{TD}}$", we convert speech represented by 25 MFCC [Imai, 1983] into 5 so-called time domain (TD) EMG features, establishing an approach parallel to the previous work on the inverse problem of generating speech from EMG signals [Janke and Diener, 2017]. In the second step, named "$\text{EMG}_{\text{TD}} \rightarrow \text{EMG}_{\text{Orig}}$", we assess the possibility of retrieving the original EMG signal from these 5 TD features.

EMG signals are speaker dependent, due to varying tissue and skin properties across speakers, as well as due to different muscle and fat proportions, and session dependent e. g., due to small shifts in the electrode positioning. For these reasons, we expect cross-session experiments to perform worse than single-session, as suggested by previous work [Diener et al., 2018]. Single-session experiments have, on the one hand, less data variability, but, on the other hand, provide a smaller amount of data for model training. Furthermore, single-session models are lacking in generalizability. Our two-step approach enables the use of simpler models in step one (session-dependent problem) and deeper models for step two (session- and channel-independent problem). The train and test partitions described in section 4.2 are maintained in all experiments. The development set was defined in each experiment, with the same dimension as the test set, as a random subset of the pool of training instances.

4.3.1 Feature extraction

Both acoustic and EMG signals were windowed using a 32 ms Blackman filter, shifted by 10 ms per step (i. e., an overlap of 22 ms). We extract 25 MFCCs [Imai, 1983] per audio frame. The choice of 25 MFCCs was motivated by the use of a similar setting in works that explore the inverse problem of EMG-to-Speech [Diener et al., 2020; Janke and Diener, 2017]. To introduce some context, we stack this feature vector with a stacking height of 15 frames into the past and future, thus representing each audio frame by a vector of dimension $25 \times 31 = 775$.

The EMG is represented as a series of 5 TD features per EMG frame: low frequency power (LF power), low frequency mean (LF mean), high frequency power (HF power), high frequency zero-crossing rate (HF ZCR), and high frequency rectified mean (HF rectified mean). The threshold for low/high frequency was 134 Hz.

This TD feature set was originally proposed by Jou et al. [2006]; Jou [2008] and has been used to convert speech-related EMG to acoustic speech in several works, such as [Diener et al., 2018], [Diener et al., 2015], and [Janke and Diener, 2017].

4.3.2 First step: $\text{Acoustic}_{\text{MFCC}} \rightarrow \text{EMG}_{\text{TD}}$

The first step was performed using a neural network containing a hourglass-shaped encoder that converts the 775-dimensional vectors representing each audio frame with context into a higher-level representation, and a set of six workers, each corresponding to one EMG channel. The encoder consists of three feed forward layers [1024, 256, 512], regularized with dropout with $p = 0.5$ and batch normalization, with rectified linear units (ReLU) activation. This hourglass shape has been employed in previous works addressing the inverse problem of converting TD EMG features to MFCCs [Diener et al., 2015]. Each of the six workers corresponds to an output linear layer with dimension 5 (each node corresponds to one TD feature) – see Figure 4.1. The loss function is based on the CCC, computed as follows:

$$loss = \frac{1}{C \times F} \sum_{c=1}^{C} (F - \sum_{f=1}^{F} CCC(y_{cf}, \hat{y}_{cf})), \tag{4.1}$$

where F is the number of features, C the number of channels, y and $\hat{y}$ are the target and predicted values. The learning rate resembles 0.002, the batch size 32, and the model was trained for 50 epochs with an Adam optimizer. As the EMG signal is session and speaker dependent, we defined three sets of experiments:

1. **Single session.** The model is trained and tested with data from the same session. We perform two single session experiments, with the two large sessions available in the corpus, which belong to speakers 2 and 8.

2. **Multi-session.** Training and testing in leave-one-session-out cross-validation setting: for each speaker, the model is trained with all training utterances of all sessions but one, and tested with the test utterances of the left-out session. We perform two multi-session experiments, with speakers 2 and 8 (speakers with a larger number of sessions).
3. **Multi-speaker.** Training and testing in leave-one-speaker-out cross-validation setting, i. e., models are trained on all training utterances of all speakers but one, and tested on all test utterances of the left-out speaker. We repeat this for the 8 speakers. In these experiments, there is no session nor speaker overlap between train and test folds.

4.3.3 Second step: $\text{EMG}_{\text{TD}} \rightarrow \text{EMG}_{\text{Orig}}$

The second step consists of synthesizing the original EMG signal from the 5 EMG TD features directly extracted from the EMG signal. Being able to retrieve the EMG signal from the TD EMG features justifies their usage as intermediary representations in future silent paralinguistics works. Furthermore, while we consider the reliable generation of EMG TD features an important proof-of-concept, the generation of the original EMG signal opens up new avenues for research and understanding of EMG signals related to spoken communication, and for silent speech, which cannot be captured acoustically.

A schematic representation of the network used can be found at the bottom of Figure 4.1. It consists of a convolutional-BLSTM neural network, similar to what has been proposed for other paralinguistic tasks, such as detection of emotions [Trigeorgis et al., 2016] and breathing patterns from speech [Schuller et al., 2020]. The neural model includes one convolutional block, one BLSTM block, and one output linear layer. The convolutional block includes three 1D convolutional layers [128, 256, 512], kernel size [5, 3, 3], stride 1, padding to keep time dimension constant, no pooling, $tanh$ as activation function, and batch normalization. The BLSTM block includes two BLSTM layers with hidden layer size 256 and dropout with a probability of 0.4, followed by batch normalization. Finally, the output linear layer has a dimension of 6 to match the sampling frequency of the EMG signal (600 Hz) and a frame shift of 10 ms used to compute the features. Each utterance is fed into the network as a tensor of dimension $f \times t$, where f is the number of TD features, and t is the number of frames in the signal. The f dimension is fed as channels for the first 1D convolutional layer, and the convolution occurs across the time dimension. The convolutional and BLSTM blocks perform a feature mapping in the sense that they keep the time dimension of the tensor constant. At the linear output layer, the time dimension is mapped to match the EMG sampling frequency. The learning rate resembles 0.001, the batch size 5, and the learning rate decay 0.1 with a period of 20 epochs. We use $1 - CCC(y, \hat{y})$ as loss function, in which y and $\hat{y}$ denote the target and predicted values. We train the model up to 50 epochs with early stopping depending on the performance on the development set.

The extraction of TD EMG features from the raw EMG signal is independent of the speaker, session,

and electrode positions (EMG channels), thus, when retrieving the raw EMG from TD features using a neural network, it is not relevant to distinguish between speakers, sessions nor electrode positions. Therefore, the data used to train and evaluate the model includes all channels from all speakers and all sessions. This results in a training set, a development set, and a test set composed of 2793, 643, and 643 utterances, respectively.

4.4 Results

CCC, used to evaluate all results, assumes values in $[-1, 1]$, where a coefficient of 0 reflects no correlation between the true and the predicted values, a coefficient of 1 reflects perfect agreement and a coefficient of -1 reflects perfect reversed agreement. CCC reflects the absolute correctness rather than only a relative one. No other metrics are reported for the sake of space and conciseness. We chose CCC over other standard metrics, such as mean squared error, because CCC assumes values in a bounded interval, easier to interpret when no previous baselines are available.

Acoustic$_{\text{MFCC}}$ $\rightarrow$ EMG$_{\text{TD}}$: session and speaker dependencies

The Acoustic$_{\text{MFCC}}$ $\rightarrow$ EMG$_{\text{TD}}$ results at the single-session (speaker 8), multi-session (speaker 8), and multi-speaker experiments are detailed in Figure 4.3. The single-session results (Figure 4.3, on the left) appear promising. We achieve a CCC value of 0.54 for speaker 2, and 0.63 for speaker 8, when averaging all feature scores across all channels. These results increase to 0.64 and 0.75 if the HF ZCR feature is excluded.

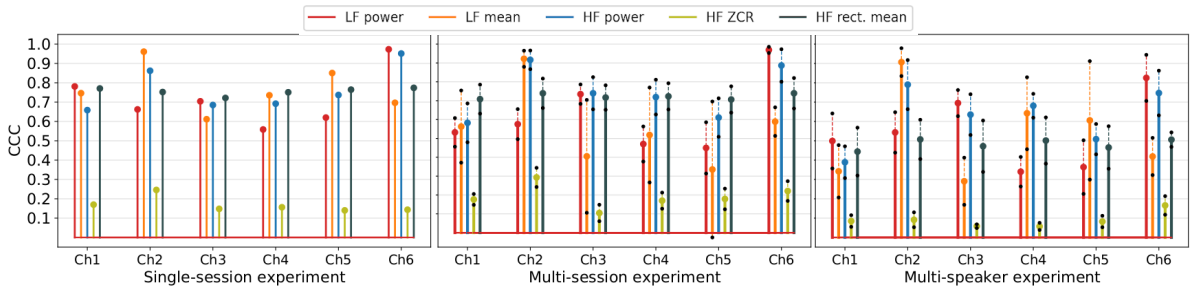


Figure 4.3: CCCs between the synthetically generated TD EMG features and the target, for the test sets of the single-session (speaker 8), multi-session (speaker 8), and multi-speaker experiments. The black dots represent the mean $\pm$ standard deviation obtained for the different sessions in the cross-validation experiments.

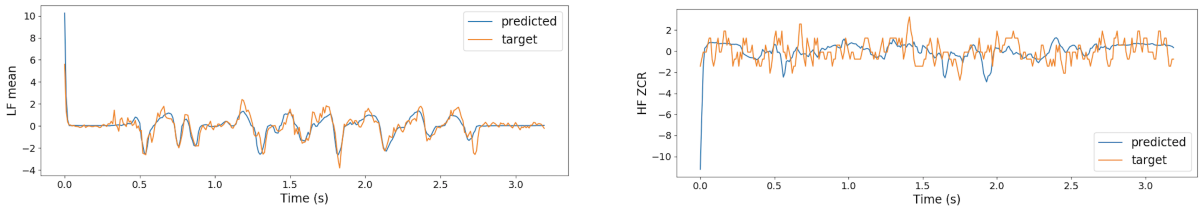
The results obtained at the multi-session and multi-speaker (Figure 4.3 – center and right) experiments appear worse when compared to the single-session experiment. The average CCC for all the features and channels for the multi-session experiments is 0.50 and 0.57, respectively, for speakers 2 and 8, while for the multi-speaker experiment, the CCC is 0.46. The average CCCs improve to 0.59, 0.66 and 0.55 when excluding HF ZCR.

The multi-session and multi-speaker experiments were evaluated in a cross validation setting. The aforementioned figure shows the mean and standard deviation of CCC for each feature at each channel, obtained for all folds. We find that there is some variation in the results for the different folds. These results support the notion that EMG signals recorded in this type of setting may be strongly session dependent, likely due to small shifts in electrode positioning. The multi-speaker results appear worse than the multi-session results, although it is not evident whether this is caused by physiological differences between speakers (e. g., fat, muscles, skin), or a result of an increasing variability in electrode position (due to an increased number of sessions, and different number of sessions per speaker) which may hinder the learning ability of the system.

$\textbf{Acoustic}_{\textbf{MFCC}} \rightarrow \textbf{EMG}_{\textbf{TD}}$: feature analysis

Figure 4.3 suggests that HF ZCR is much harder to predict than the rest of the TD features in all the experiments. Figure 4.4 presents an example of the target and the predicted LF mean and HF ZCR of channel 0, obtained in the single-session experiment with speaker 8, to illustrate the meaning of the different CCCs.

The CCCs achieved at the single session experiments for LF mean, LF power, HF rectified mean, and HF power are above 0.5 for all channels. These results are at a comparable level with the prediction of the first three MFCCs when converting audible EMG to acoustic speech [Diener et al., 2020]. For the remaining MFCCs, the results in the speech-to-EMG direction are much better than the results in the direction EMG-to-speech.



(a) LF mean. CCC $= 0.870$ **(b)** HF ZCR. CCC $= 0.182$

Figure 4.4: Examples of the target and predicted features LF mean and HF ZCR (single-session, speaker 8).

$\textbf{EMG}_{\textbf{TD}} \rightarrow \textbf{EMG}_{\textbf{Orig}}$

We use the five target TD features to generate the original EMG signal. Figure 4.5 shows an example of the predicted and target EMG signals, suggesting a reasonable match.

As HF ZCR appeared to be harder to predict than the remaining features, we also generate the speech-related EMG signal based on the remaining four TD features. Table 4.2 shows that although the CCCs obtained with both feature sets are very satisfactory, the results are better in the presence of ZCR. Thus, we conclude that ZCR contains relevant information for the generation of EMG_{Orig}.

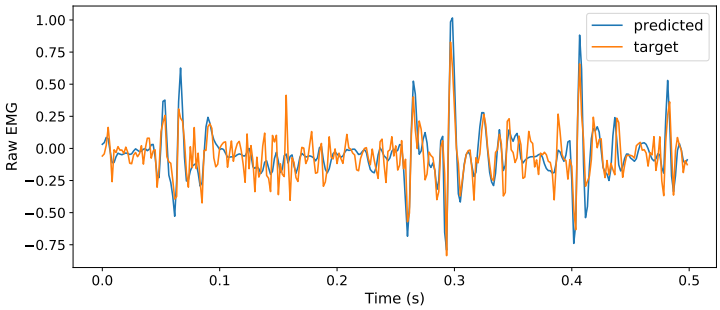


Figure 4.5: EMG signal generated from TD features.

Table 4.2: CCC between the target and the predicted EMG signal, using four and five TD features.

Input Features	CCC
LF mean, LF power, HF rectified mean, HF power, HF ZCR	0.663
LF mean, LF power, HF rectified mean, HF power	0.602

4.5 Summary

The results described in this chapter, together with the results present in [Diener et al., 2020] established the foundation for the novel field of Silent Computational Paralinguistics, which opens up new avenues for research with EMG signals related to spoken communication, for a deeper understanding of speech production, and for silent speech interfaces.

In this chapter, we presented initial results on a novel two-step approach to generate speech-related EMG signals from acoustic speech. In the first step, we successfully converted MFCCs into TD EMG features. Thus, we established the foundation of the speech-to-EMG approach - i. e., the inverse of prior works that have aimed to generate speech from EMG data. The CCCs achieved in the single session experiments for the prediction of LF mean, LF power, HF rectified mean, and HF power were comparable to the first three MFCCs generated from audible EMG in our work on EMG-based Silent Computational Paralinguistics [Diener et al., 2020]. Multi-session and multi-speaker experiments, although performing worse than the single-session experiments, still achieved satisfactory results. We expect that this may be improved with deeper and more complex models when more data is available. In the second step, we generated a signal that follows reasonably the true EMG signal, using the TD features.

It is possible that an end-to-end approach using an encoder-decoder architecture may achieve better results. However, our early experiments on an end-to-end setting did not achieve satisfactory results, possibly due to data scarcity. We leave for future work this integration of both speech-to-EMG and EMG-to-speech in one system, as well as the retrieval of paralinguistic information from the generated EMG signals.

We anticipate that with the acquisition of additional data, the emerging field of Silent Computational Paralinguistics will play a role in disease detection. By investigating biosignals beyond acoustics, it may

be possible to differentiate diseases based on which parts of the speech production process are most disrupted. However, further data collection was not possible due to the COVID-19 pandemic, hindering this line of research. Consequently, in the following chapters, we redirected our focus to other biomarkers that can be collected remotely.

5

Using speech and complementary modalities for disease detection: the case of obstructive sleep apnea

Contents

THIS chapter builds upon research conducted during my Master's Thesis, which focused on speech-based detection of sleep deprivation and sleep disorders, specifically obstructive sleep apnea (OSA), in two scenarios: controlled data collection in Portuguese and in-the-wild data. Here, in addition to analyzing the speech signal, we investigate facial images and visual speech as three complementary modalities for automatically detecting OSA. While our primary focus is on OSA, we anticipate that these modalities may also prove valuable for detecting other speech-affecting diseases.

Our experiments compare knowledge-based features and transfer learning methodologies in a pilot in-the-wild corpus with 40 subjects. We further discuss the suitability of using in-the-wild data from Youtube vlogs for the automatic detection of diseases. The work described in this chapter has been published at Interspeech 2021 [Botelho et al., 2021].

5.1 Introduction

As previously described in section 2.2.1, it is estimated that approximately one-seventh of the world's adult population suffers from OSA [Benjafield et al., 2019], and the numbers tend to increase with the growth of OSA's main risk factors: obesity and population aging. Thus, more comfortable, cost-effective and less time-consuming alternatives for OSA's diagnosis are required. In particular, biomarkers are on demand, which could be collected non-invasively, and remotely. Together with technical systems that support medical diagnoses, such automatic OSA detection would allow for large scale screenings in the patients' homes.

Several authors have proposed systems for automatically detecting OSA, using machine learning algorithms fed by biomarkers derived from speech [Benavides et al., 2014; Botelho et al., 2019; Kriboy et al., 2014a; Perero-Codosero et al., 2019] and facial images [Balaei et al., 2017; Lee et al., 2009a; Nosrati et al., 2016]. The success of these biomarkers builds on the fact that OSA patients have anatomical and functional abnormalities of the upper airway and an altered craniofacial morphology, which impacts both speech and facial expressions, as previously described in section 2.2.1.

Our approach consists of tackling OSA detection using not only acoustic speech, and facial images, but also a third modality: visual speech. There is a long history of research on visual speech recognition [Chung and Zisserman, 2016; Ma et al., 2021; Zhou et al., 2014], and a growing interest in audio-visual speech enhancement and separation [Gabbay et al., 2017; Tan et al., 2020]. Nevertheless, to the best of our knowledge, visual speech has not yet been explored in the context of Silent Computational Paralinguistics. We argue that embeddings trained for lip reading also encode information on the craniofacial structure, speech articulation and breathing patterns. For the particular problem of OSA detection, which has been shown to benefit from information derived from both speech and craniofacial structure, we hypothesize that visual speech may conjugate both domains. Furthermore, this modality may be robust

when using in-the-wild data, in which the speech signal is often contaminated with music, noise and other voices.

Large health-related data sets are difficult to acquire due to time and monetary constraints, lack of awareness, as well as ethical, legal, social, and privacy concerns. Thus, in this work, we decided to follow the idea proposed by Correia et al. [2018b, 2021], and collected a corpus of in-the-wild data, composed of YouTube vlogs of 40 subjects, roughly half of which claimed to suffer from OSA. In fact, this corpus corresponds to an extension of the WOSA corpus, a small pilot corpus we collected in earlier work [Botelho, 2018; Botelho et al., 2019]. Although we frame this problem with in-the-wild data, which is easier to acquire in larger scale, we expect that the three modalities can be of great relevance in the context of online medical appointments or remote population screenings where both audio and visual data is available. Acknowledging concerns surrounding the use of in-the-wild data, Appendix B details a series of experiments where classifiers trained on in-the-wild vlogs for detecting Parkinson's disease and depression are tested on medically verified data collected under controlled conditions. These experiments demonstrate promising results, suggesting that in-the-wild vlogs are suitable for disease detection and serve as an initial step when other data sources are unavailable. These findings are presented in the Appendix to maintain focus within this chapter, as they use different corpora and concentrate solely on the speech modality.

5.2 Related Work

The altered craniofacial features characteristic of individuals who suffer from OSA, which were thoroughly described in section 2.2.1, motivated several researchers to use facial images and speech signals for OSA detection. These works have mostly proposed systems based on classic machine learning methods and knowledge-based features. We note that the related work discussed here reflects the existing literature available at the time these experiments were conducted.

5.2.1 OSA detection from speech signals

Given the articulatory, phonation and resonance anomalies expected to be present in the speech of OSA patients (see section 2.2.1), several authors have addressed the automatic detection of OSA through voice analysis, using corpora in Spanish [Benavides et al., 2014; Espinoza-Cuadros et al., 2016; Perero-Codosero et al., 2019; Pozo et al., 2009; Solé-Casals et al., 2014] and Hebrew [Elisha et al., 2012; Goldshtein et al., 2011; Kriboy et al., 2014a,b]. The most common acoustic features in these works are Mel frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC), Energy, Harmonics-to-noise ratio (HNR), jitter, and formants frequency and bandwidth. The most common classification and regression methods are Gaussian Mixture Models (GMM) [Elisha et al., 2012; Goldshtein

et al., 2011; Pozo et al., 2009], Linear Discriminant Analysis (LDA) [Benavides et al., 2014; Kriboy et al., 2014a], k-Nearest Neighbors (kNN) [Kriboy et al., 2014a; Solé-Casals et al., 2014], Support Vector Machines (SVM) [Solé-Casals et al., 2014], Bayesian Classifiers [Solé-Casals et al., 2014], Neural networks [Solé-Casals et al., 2014], Adaboost [Solé-Casals et al., 2014], and Support Vector Regression (SVR) [Espinoza-Cuadros et al., 2016; Kriboy et al., 2014b].

Goldshtein et al. [2011] made a separate analysis for female and male speakers, and they reported better results for female speakers. A limitation of their work is the fact that they manually segmented the vowels and the two nasal phonemes, /n/,/m/, from speech signals. The work of Elisha et al. [2012] followed the work of Goldshtein et al. [2011], and it concluded that the phonemes carrying more distinguishing information were the vowel /a/ and the nasal phonemes (/m/ and /n/), which is consistent with the resonance anomalies previously described in section 2.2.1.

Espinoza-Cuadros et al. [2016] collected the largest corpus, having obtained worse results than the other works, in both regression and classification tasks. This motivated them to make a careful review of previous works and on possible pitfalls that could be responsible for overoptimistic results. The authors pointed out three main pitfalls: small and very often unbalanced corpora, which are more prone to overfitting; presence of confounding variables such as gender, age and body mass index unevenly distributed between classes; and feature selection on high dimensionality feature spaces when little data is available, which is also most likely to cause data overfitting.

The works of Kriboy et al. [2014a] and Solé-Casals et al. [2014] hypothesized that acoustic properties of speech that are altered by body position help distinguish between OSA and non-OSA subjects. In fact, there is an increased frequency and severity of apneas in supine position, most likely due to unfavorable airway geometry, increase in collapsibility, gravity and inadequate dilator muscle compensation.

Fewer works have leveraged deep learning methodologies for OSA detection. An exception is the work proposed by Perero-Codosero et al. [2019] which explores *x-vectors* and domain adversarial training for OSA detection from speech.

5.2.2 OSA detection from facial images

Lee et al. [2009a] studied the craniofacial morphological phenotype of subjects with and without OSA using a quantitative photographic analysis technique. To this end, they used frontal and profile standardized images of the subjects' face, where a set of landmarks was manually annotated (see Figure 5.1). Afterwards, they derived a total of 71 measurements of various craniofacial regions. They concluded that it was possible to identify craniofacial phenotypic differences in OSA in Caucasian subjects. In this study, Lee et al. used a dataset including 180 subjects, of which 114 had OSA (respiratory events $>=$10/h) and 66 were selected as controls (number of respiratory events $<$10/h).

Nosrati et al. [2016] proposed to use the same 71 craniofacial measures to automatically detect

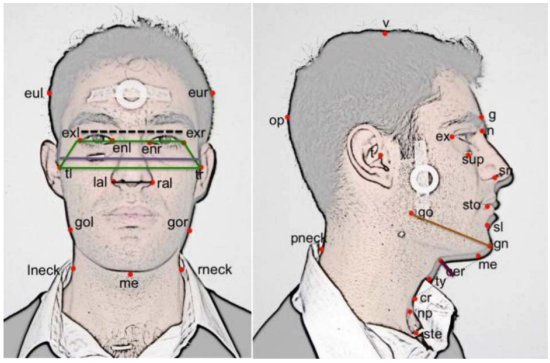


Figure 5.1: Photographic landmarks used in [Balaei et al., 2017; Lee et al., 2009a; Nosrati et al., 2016]. Image from [Nosrati et al., 2016].

whether a subject suffered from OSA. They used frontal and profile photographs of the same 180 subjects. A calibration object of known size was taped to the forehead and the cheek of the subjects, before taking the photographs. It was used to calibrate measurements determined from the photograph. Nosrati et al. found that, out of the 71 measurements, the most informative features were a set of five features, including face width *(tl-tr)*, mandibular length 2 *(gn-go)*, binocular width *(exl-exr)*, cranial base area *(tl-exl-exr-tr)*, and criocomental space distance *(cer-cr-me)*. Using these five features and an SVM trained in a leave-one-out cross-validation training fashion, they achieved an accuracy of 73.3% for the binary detection of OSA (chance level: $\approx$ 63%).

One key limitation of the aforementioned works is the reliance on manual landmark annotation. Balaei et al. [2017] proposed to overcome this limitation, by training a system that automatically identifies the same facial landmarks. Afterwards, they derive four measurements or features, common to the set of 71 features proposed by Lee et al. [2009b]: face width *(tl-tr)*, eye width *(exl-exr)*, cervicomental angle *(np-cer-me)* and mandibular length 1 *(me-go)*. Two out of these four measurements also belong to the set of features identified as most relevant by Nosrati et al. [2016]. In their work, Balaei et al. studied 365 subjects, 223 with OSA and 142 control subjects. Out of the 365, 169 subjects are common to Lee's dataset [Lee et al., 2009b]. The calibration of the new subjects' photographs was performed using a laser ruler with known size, projected onto the face for the photograph. They achieved an accuracy of 69.8% for binary OSA detection using the automatic landmarks to extract the four features, followed by a logistic classifier (chance level $\approx$ 61%). Furthermore, they observed that the automatic landmark detection introduced approximately a 10% error relative in the determination of the features. Balaei et al. also experimented an alternative approach, that feeds directly the downsampled images to a feed forward neural network, without the step of calculating facial landmarks and features. With this approach, they achieve 61.8% accuracy (chance level is not reported, although it is mentioned that the test set is chosen randomly from the data subset used for this experiment, which contains 61% of OSA subjects).

In another study, Balaei et al. [2018] compared not only the manually and automatically extracted

landmarks followed by feature computation, but also the feasibility of using directly the aligned landmarks' coordinates as input to the classification method. Their study, which used frontal and profile facial photographs of 376 subjects, showed that using directly the landmarks' coordinates, without the feature computation step, achieved better results. With this approach, they achieved an accuracy of 69.7% with manually annotated landmarks, and 69.2% with automatically extracted landmarks (the ratio of OSA and controls in the test set is not reported, although it is 62:38 in the entire dataset).

Espinoza-Cuadros et al. [2015] have addressed the estimation of OSA's severity, using uncalibrated craniofacial features, computed from automatically detected landmarks. They compared the craniofacial features to speech signals, represented by *i-vectors*, and observed that craniofacial features were better predictors of OSA severity than *i-vectors*. In fact, they point to a weak connection between OSA and speech. In their work, they studied a dataset of 285 male subjects and, although their main task was regression, they also provide results for the binary classification, achieving an accuracy of 70.8% (chance level was not reported).

5.2.3 Visual Speech

In this work, we explore a third modality, inspired by the recent progress in audio visual recognition, particularly for in the wild datasets. This progress has been shown for the Lip Reading in the Wild (LRW) Dataset [Chung and Zisserman, 2016], which comprises video clips from over 1000 speakers from BBC programs, and includes a vocabulary of 500 English words. Each video clip is one second long, and contains the target word surrounded by other context words. The lip reading task is then framed as a multi-class classification problem, which predicts, for each video clip, the spoken word. Many approaches have been proposed to tackle this challenging task. One of such approaches is the ensemble of models proposed by Ma et al. [2021]. In their work, each model in the ensemble consists of a modified ResNet-18 which leverages born-again distillation (iterative self-distillation) for improving the performance. Their best single model corresponds to the third generation of knowledge distillation, and achieves a top-1 accuracy of 87.9% on the LRW dataset. The ensemble of methods achieved a state of the art top-1 accuracy of 88.5%, a promising value for the use of visual speech as an extra modality in paralinguistic tasks.

5.3 Method

We address OSA automatic detection using three modalities – acoustic speech, facial images, and visual speech –, extracted from vlogs collected from YouTube. For each modality, we start by a data selection and pre-processing stage, followed by feature extraction and binary classification. We also perform early and late fusion of the three modalities.

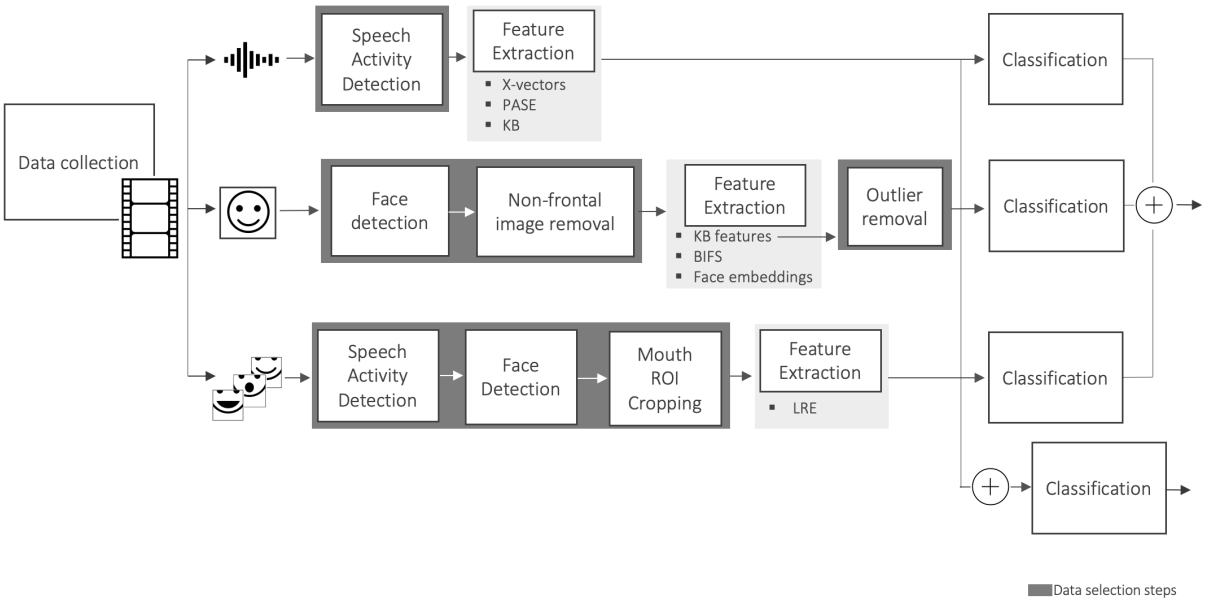


Figure 5.2: Methodology pipeline.

We compare both knowledge-based features and neural representations obtained using transfer learning from different but related tasks, for which more data was available. In particular, one of the neural representations used were embeddings trained for identity recognition for both speech and facial image modalities.

The binary classification step was performed using a neural network (NN) classifier, trained in a leave-one-subject-out cross-validation setting to work around the limited number of subjects.

A schematic representation of the methodology is depicted in the Figure 5.2. Below, we describe in detail the data collection procedure, the features and neural network architectures used for each of the modalities. The descriptions related to the speech modality are more succinct, as most of the techniques have been thoroughly described in chapter 3.

5.3.1 Data collection

Following the idea presented by Correia et al. [2018b, 2021], we collected vlogs publicly available on YouTube, where subjects claim to have OSA. The platform was queried using the keywords “obstructive sleep apnea vlog”, “sleep apnea vlog”, “my obstructive sleep apnea”, and “my CPAP[1] review”. Before selecting the videos, we manually verified that subjects claimed to suffer from OSA or weight-related sleep apnea and that it was not a different kind of video, such as lectures on OSA, medical professionals explaining OSA, or advertisement on CPAP machines, where the main subject does not claim to suffer

[1] CPAP is the most efficacious and widely used treatment option for obstructive sleep apnea. It is a ventilator that applies a mild pressure on a continuous basis, through a mask that the patients wear during the night. The positive airway pressure prevents the airways to collapse or to become blocked making the patient unable to breath [Arnold et al., 2017].

from OSA. After selecting the videos of subjects claiming to have OSA, we used a subset of the control videos that are part of the WSM corpus [Correia et al., 2021]. These videos were also collected from YouTube, and were queried using unrelated keywords, such as "book review", "lets talk vlog", "lets talk knitting", to serve as control subjects. The selection was carefully performed in order to collect the same number of videos featuring male and female subjects, both for OSA and control classes.

The resulting corpus corresponds to a second version of the WOSA corpus, which we have collected in earlier work [Botelho, 2018; Botelho et al., 2019].

While we are aware that the labels in this dataset are noisy and not medically verified, our approach is motivated by the success of prior work. For example, in [Botelho et al., 2019] we successfully used speech recordings of similar datasets for OSA classification, and [Correia et al., 2018b, 2021] applied the approach to the detection of Parkinson's disease and depression (WSM corpus). We have also performed a cross-corpus comparison, where we trained models using the WSM corpus, i.e. using in-the-wild data, and tested with data from controlled conditions with medically verified labels. These results, presented and discussed in Appendix B, are reassuring as they present evidences for the suitability of in-the-wild data.

5.3.2 OSA detection using acoustic speech

To address the detection of OSA from acoustic speech, we run a speech activity detector, followed by feature extraction, and binary classification, as illustrated on the top branch of Figure 5.2.

Data cleaning and pre-processing

The original audio files were converted to mono and downsampled to 16kHz. Afterwards, we experimented using a Python interface of the WebRTC Voice Activity Detector (VAD) [Wiseman, retrieved in March 2021], but we observed that it classified as speech several segments with music, noise, and/or sound effects. We opted to perform speech/non-speech segmentation using an in-house VAD which consists of a feed-forward NN trained with perceptual linear prediction (PLP) features, followed by a finite state machine, trained with broadcast news [Meinedo and Neto, 2005]. Nevertheless, the segmentation was not free or errors and some music/sound effects were still introduced as speech segments in the analysis.

Feature extraction

We compared three types of features: *x-vectors*, embeddings extracted with the *PASE+*, and *knowledge-based* features. The architecture and intuition for using X-vectors and PASE+ embeddings in paralinguistic tasks has been described in chapter 3.

X-vectors were extracted using Kaldi [Povey et al., 2011], with a model pre-trained on VoxCeleb [Nagrani et al., 2019]. VoxCeleb, previously described in section 3.3, contains clips of celebrity interviews

on YouTube. Although the quality of interviews in VoxCeleb is probably better than those of the "home-made" vlogs included in the corpus we collected for OSA detection, the domain is arguably similar. Prior to *x-vector* extraction, we extracted 30 MFCCs computed every 10 ms from 25 ms-length frames. We applied again VAD to filter out remaining non-speech frames, and perform cepstral mean and variance normalization. The resulting x-vector embeddings were 512 dimensional.

The *PASE+* embeddings were extracted using a pre-trained model[2] [Pascual et al., 2019; Ravanelli et al., 2020]. Each audio segment was thus represented by a matrix with the dimension $256 \times n\_frames$.

The *knowledge-based (KB)* feature set corresponds to 109 features, proposed in [Botelho et al., 2019] for OSA detection. This feature set includes features related to formants, harmonics-to-noise ratio, jitter, F0, Spectral Flux, MFCCs plus their first and second order derivatives (ΔMFCC and $\Delta\Delta$MFCC), and LPCC.

Classification Experiments

We defined three experiments to perform OSA classification using speech signals:

A. X-vectors Experiment: The *x-vectors* were fed to a fully connected feed forward neural network.

B. PASE+ Experiment: The PASE+ embeddings were fed to a convolutional neural network (CNN).

C. KB features Experiment: The KB features were fed to a fully connected feed forward neural network.

The choice of the neural architecture depended on the input type: x-vectors and KB features represent each audio segment with a fixed size vector, and thus were fed to a fully connected feed forward neural network; PASE+ features have a dimension that depends on the duration of the audio input, and thus were fed to a 1D CNN, followed by a statistical pooling layer. All three neural networks were trained using cross entropy loss, in which each class was weighted by the inverse of its relative frequency in the training folds. Appendix C provides further details on the networks' architecture and training hyperparameters.

5.3.3 OSA detection using facial images

The overall system that uses facial images for OSA detection, as depicted in the middle branch of Figure 5.2, consists of five main steps for data cleaning and processing, feature extraction and finally a binary classification component based on neural networks to perform the OSA detection.

Data cleaning and pre-processing

We started by extracting the key frames of each video. Given that the videos were obtained in-the-wild, we have no control over the subject's face, at which angle it appears, if it is occluded or if it appears at all in the key frames. Thus, a pre-processing step removed all frames unsuitable for the classification task.

[2] https://github.com/santi-pdp/pase

The pre-processing included face detection, non-frontal image removal, and removal of outliers. After pre-processing, we made sure that all the subjects included in the analysis had at least 5 facial images.

For data cleaning and feature extraction different pre-trained models were used, thus images had to be resized and cropped in different ways. For face detection, images were resized to (300, 300) to match the pre-trained model. If the original dimension was not square, the smallest dimension was padded with zeros. After face detection, all images were cropped around the face, using the face box identified by a face detector (see below for further details). If the face box had the dimensions (h, w), the cropping area is a square box, with dimensions $h_{cropping} = w_{cropping} = max(h + \alpha h, w + \alpha w)$, where α represents a margin around the face box, to avoid cropping the chin, forehead and neck. α was set to 1.

Face detection. The first step to selecting frames relevant for OSA detection is face detection. For this step, we used the pre-trained deep neural network (DNN) face detector in *OpenCV*, which is based on the single shot detector (SSD) [Liu et al., 2016], using a ResNet-10 network. The SSD is a deep learning approach for object detection. It is based on a CNN that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes, followed by a non-maximum suppression step to produce the final detections [Liu et al., 2016]. The model was trained using images available from the web, but the source was not disclosed. The pre-trained *OpenCV* DNN model and configuration for face detection are publicly available[3]. We used the 8-bit quantized version using *Tensorflow*. The minimum confidence for accepting the face as detected was set to 0.98. All frames with multiple or zero faces detected were excluded.

Non-frontal image removal. To assess if a face is approximately in frontal position, we computed 68 facial landmarks (Figure 5.3) using *Dlib*'s pre-trained model[4]. The model consists of an ensemble of regression trees learnt via gradient boosting, which can be used to estimate in real time the face's landmark positions directly from a sparse subset of pixel intensities. The model's implementation was based on [Kazemi and Sullivan, 2014], and it was trained on the iBUG 300-W face landmark dataset [Sagonas et al., 2016]. Based on the computed landmarks, a face is considered to be in frontal position if (i) the distance of each eye to the face margin is approximately the same for both eyes; and (ii) the distance of the rightmost landmark of the mouth to the margin of the face is approximately the same as the distance of the leftmost landmark of the mouth to the margin of the face. Conditions (i) and (ii) are expressed in equations 5.1 and 5.2, respectively,

$$min\left(\frac{dist(l_{37}, l_1)}{dist(l_{46}, l_{17})}, \frac{dist(l_{46}, l_{17})}{dist(l_{37}, l_1)}\right) > 1 - \tau \tag{5.1}$$

[3] https://github.com/spmallick/learnopencv/tree/master/AgeGender
[4] http://dlib.net/files/shape_predictor_68_face_landmarks.dat.bz2

$$min\left(\frac{dist(l_{49}, l_4)}{dist(l_{55}, l_{14})}, \frac{dist(l_{55}, l_{14})}{dist(l_{49}, l_4)}\right) > 1 - \tau \tag{5.2}$$

where $dist(l_i, l_j)$ represents the Euclidean distance between landmarks l_i and l_j, l_i corresponds to the i^{th} landmark in Figure 5.3, and τ is a tolerance term set to 0.1, to account for possible face asymmetries, or minor landmark computation errors.

Feature Extraction

We compared three sets of features for OSA detection: KB, bio-inspired features (BIF) features, and face embeddings. Prior to computing BIF, the face images were converted to black and white and resized to (100, 100), to obtain vectors with constant dimensions.

Knowledge-based features. Following [Balaei et al., 2017; Espinoza-Cuadros et al., 2015; Nosrati et al., 2016], we defined a set of five KB features: face width, binocular width, mandibular length, cranial base area, and mandibular nasation angle, described in Figure 5.3 and Table 5.1.

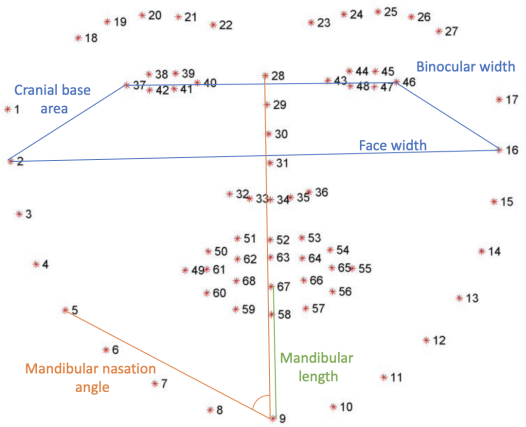


Figure 5.3: KB features.

Auxiliary measures	
Eye width	$euclidean\_distance(l_{37}, l_{40})$
Eye area	$area(l_{37}, l_{38}, l_{39}, l_{40}, l_{41}, l_{32})$
Features	
Face width	$\frac{euclidean\_distance(l_2, l_{16})}{eye\_width}$
Binocular width	$\frac{euclidean\_distance(l_{37}, l_{46})}{eye\_width}$
Mandibular length	$\frac{euclidean\_distance(l_{67}, l_9)}{eye\_width}$
Cranial base area	$\frac{area(l_{37}, l_{46}, l_2, l_{16})}{eye\_area}$
Mandibular nasation angle	$angle(l_9, l_{28}, l_5)$

Table 5.1: KB features defined in terms of landmarks l_i.

The KB features used in this work do not match exactly those of previous works (figure 5.1), and the definition had to be adjusted, for four reasons: (i) we do not have access to profile images of the main subject; (ii) we use the landmarks extracted as described in section 5.3.3 to compute the features, which are different from those used in previous works; (iii) the distance of the subjects to the camera varies substantially, both within and across vlogs; and (iv) there is no object of known size to allow for calibration. To deal with (iii) and (iv), we made the assumption that individuals have roughly the same eye width and area, and used these measures to normalize all other features. Although this is a coarse approximation, it allows to calibrate the measures that are going to be used as input features for the classification system.

Bio-inspired features. BIF, proposed by Guo et al. [2009] for age estimation, are extracted in two steps

or layers: layer S_1 and layer C_1. At layer S_1, the cropped face is filtered by a set of Gabor functions at different orientations and scales. Afterwards, layer C_1 performs a pooling operation which keeps the maximum and the standard deviation of the Gabor filtered outputs. Previous bio-inspired models only use the maximum for the pooling operation. Guo et al. introduced the standard deviation in the pooling operation because it captures local variations which might be important to characterize the subtlety of aging (e.g., wrinkles, creases, and eyelid bags). Although these subtleties may be crucial for age estimation, it is not clear whether they will be an advantage for OSA detection, given that OSA has been mainly associated with larger altered craniofacial morphology. Nonetheless, there is evidence that short term [Kim et al., 2017] and long term [Jang et al., 2020] sleep deprivation induces skin alterations, such as decreased skin hydration, decreased skin elasticity, decreased skin blood flow and altered skin pores, which could be captured by BIF and thus contribute for OSA detection from facial images. On the other hand, considering that our dataset was collected from YouTube vlogs, it is possible that these skin alterations are eliminated on purpose by the subjects, through makeup or image processing techniques to soften the appearance of their skin. We used *OpenCV*'s implementation of BIF[5], using the default number of bands (8) and rotations (12).

Face embeddings for face recognition. The facial embeddings were extracted using *Dlib*'s pre-trained model for face recognition[6] [King, 2017], which consists of a version of the ResNet-34 described in [He et al., 2016], with 29 convolutional layers. The model was trained on a dataset of about 3 million faces and 7485 individuals, derived from the face scrub dataset [Ng and Winkler, 2014], VGG dataset [Parkhi et al., 2015] and other images scraped from the internet. The NN was evaluated on the standard Labeled Faces in the Wild benchmark [Huang et al., 2007], achieving 99.38% accuracy. The embeddings generated are vectors of size 128. Although face recognition and disease diagnosis are different tasks, one can hypothesize that they are related enough to allow the transfer of meaningful information. As previously described, the craniofacial morphology of the subject contains information useful for OSA detection, and this same craniofacial morphology is likely to be a key-component for face recognition.

Outlier removal

At this point, all frames that do not include approximately frontal face images were automatically removed. However, there are still frames in the pool that may show a non-target subject or any object mistakenly classified as face. To remove these frames, we performed outlier removal based on interquartile range (IQR) scores. IQR is a measure of statistical dispersion, being equal to the difference between 75^{th} (Q3) and 25^{th} (Q1) percentiles, $IQR = Q3 - Q1$. So, a data point p is considered to remain in the pool if inequation 5.3 is satisfied. To define whether a face image is an outlier, we represented each face in the image by the KB features.

[5] https://docs.opencv.org/3.4/dc/d12/classcv_1_1face_1_1BIF.html
[6] https://github.com/davisking/dlib-models?tab=readme-ov-file#dlib_face_recognition_resnet_model_v1datbz2

$$\begin{cases} p > Q1 - 1.5 * IQR \\ p < Q3 + 1.5 * IQR \end{cases} \tag{5.3}$$

Classification Experiments

We defined six classification experiments to perform OSA detection based on the raw images, and on each of the feature sets described above, using different NNs:

A. Facial images Experiment: In this experiment, the raw images, previously resized to (100, 100), were fed to a CNN.

B. Facial images Experiment with local attention: Experiment A was repeated, adding an attention block. The idea of local attention is introduced by the convolutional layer in the attention block, which filters the inputs and selects which pixels to give more weight. This experiment was designed to provide some explanation on which part of the image was considered more relevant by the network, for the classification task.

C. Facial images Experiment with global attention: Experiment A was also repeated, using global attention. The global attention layer follows the standard attention layer architecture.

D. Knowledge-based features Experiment: The KB features were fed to a fully connected NN.

E. Bio-inspired features Experiment: The BIF vectors were fed to a fully connected NN.

F. Facial embeddings Experiment: The embeddings were fed to a fully connected NN.

Appendix C provides further details on the network's architectures and training hyperparameters.

5.3.4 OSA detection using visual speech

We address the detection of OSA from visual speech as depicted in the lower branch of Figure 5.2. We started by a data selection step that includes VAD and face detection to ensure that we only attempt OSA detection in segments where the main subject is speaking while appearing on camera. Then we extract lip reading embeddings which are fed to a NN, for classification.

Data cleaning and pre-processing

First we performed VAD, using the Python interface of WebRTC VAD, previously mentioned in section 5.3.2. Secondly, we took the audio chunks for which voice was detected and split those into one-second segments. Unlike the methodology used in [Chung and Zisserman, 2016], we did not attempt to perform a segmentation which includes a given target word, because we were not aiming at classifying any particular words. Instead, our goal was to capture articulation and breathing patterns together with the craniofacial morphology, which can encode paralinguistic information. Thus, we segmented the

audio chunks resultant from VAD sequentially, into one-second segments.

Afterwards, we performed face detection on all the frames that correspond to the one-second voice clips. For this step, we used the same model used for the facial images methodology, described in section 5.3.3. If the face detector did not find a face in at least 75% of the frames that correspond to a one-second clip, that clip was excluded. Then, for the faces in the selected clips, we computed the 68 facial landmarks, using the same model described in 5.3.3. The faces were aligned to a reference frame, following the methodology in [Ma et al., 2021]. Finally, we cropped the faces around the mouth region (region of interest) to a bounding box 96×96, following [Ma et al., 2021].

Feature extraction

The lip reading embeddings used here were extracted using a pre-trained model[7] proposed by [Ma et al., 2021] (see section 5.2.3). As mentioned before, Ma and Martinez's best single model corresponds to the third generation of knowledge distillation. We used this third generation ResNet-18 to extract the lip reading embeddings. Each clip is represented by a matrix of size $n\_frames \times 512$. The vlogs have slightly different frame rates, thus, we use the first 25 frames of each clip, which is approximately one-second.

Classification Experiments

The 25×512 lip reading embeddings were fed to a CNN. Further details on the network architecture and training hyperparameters are described in Appendix C.

5.3.5 Fusion of the three modalities

We compare two strategies to combine the three modalities: late fusion (experiment A) and early fusion (experiment B).

A. Late fusion experiment: After making a prediction for each subject, for each modality, we performed a majority vote to assign a final prediction to each subject. We used the predictions of the best performing system of each modality.

B. Early fusion NN experiment: To perform early fusion, we used the facial embeddings to represent the image modality, the x-vectors for the speech modality, and the lip reading embeddings for the visual speech modality. Because the pre-processing of the three modalities was done independently from each other, the segments are not synchronous in time, and their quantity is different across the three modalities. Thus, we randomly sampled 100 instances of each modality per subject, with replacement. Afterwards, we fed the three embeddings to a NN that follows the structure represented in Figure 5.4. The neural network architecture and training hyperparameters are further detailed in Appendix C. We repeated the sampling process 100 times to create 100 distinct datasets. Then, we trained 100 models,

[7] https://sites.google.com/view/audiovisual-speech-recognition

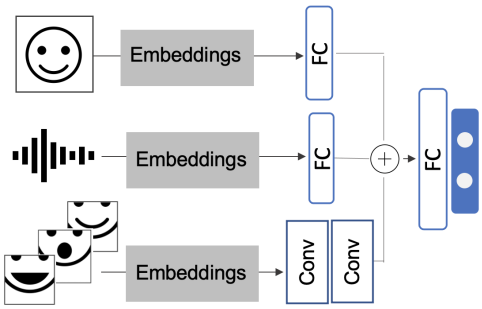


Figure 5.4: Network architectures for the early fusion experiment B. *FC* stands for fully connected block, and *Conv* stands for convolutional block.

which allows us to present the results with a 95% confidence interval.

5.4 The WOSA-2.0 corpus

The unprocessed dataset initially contained YouTube vlogs of 47 subjects. After processing within the facial images modality, 7 subjects were excluded, resulting in a final corpus of 40 vlogs publicly available on YouTube. This corpus, named WOSA-2.0, is an extension of our previous work [Botelho, 2018; Botelho et al., 2019]. Not all 16 speakers from the original WOSA are included here, as some videos were removed from YouTube in the interim. Of the 40 vlogs, 22 feature subjects who claim to suffer from OSA, while the remaining 18 serve as control subjects. The gender distribution is presented in Table 5.2.

(a) Control subjects **(b)** OSA patients

Figure 5.5: Examples of images included in the final dataset for analysis.

Figure 5.5 shows examples of facial images included in the corpus, to illustrate the control and OSA vlogs that were included in the corpus. Although most images are nicely cropped and in frontal position, there are a few images where the subjects are not exactly on frontal position. Moreover, there are subjects wearing sunglasses or with eyes almost closed. Thus, it is likely that the KB features in the facial images modality are affected by this variability.

Table 5.2: Corpus participants.

	OSA subjects	Control subjects
Total	22	18
Male	12	9
Female	10	9

Table 5.3: Corpus: instance counts and duration per modality.

	Total count	Count per subject mean ± std	Duration (s) mean ± std
Facial images	2733	68 ± 56	–
Audio files	4953	124 ± 133	4 ± 7
Video clips	22261	557 ± 401	1 ± 0

Although the subjects claim to suffer from OSA and talk about their disease, we believe that this fact does not compromise the results using lip reading embeddings. We manually verified that the vocabulary of the Lip Readind in the Wild (LRW) corpus, used for training the lip reading extractor, does not contain the target words "obstructive sleep apnea", "apnea", "sleep", "disease", nor "disorder". The only words in the vocabulary related to healthcare were "hospital" and "medical". For the vlogs for which the automatic transcription was publicly available (38 out of 40) we counted the number of occurrences of those two words. The word hospital occurs 5 times (0.02%), all of them spoken by subjects with OSA, and the word "medical" occurs 35 times (0.16%), 32 of which spoken by subjects with OSA. This assumption could be further minimized, in future work, by considering control data focused on medical vlogs, and thus ensuring a more similar vocabulary. Besides the vocabulary usage, it is possible that the paralinguistic content, namely the emotional tone, of the vlogs where the main subjects are describing their own experience with the disease differ from those where controls subjects are discussing unrelated control topics, or even medical vlogs. This limitation is addressed by the experiments in Appendix B.

The different pre-processing steps for each modality described in section 5.4 resulted in different numbers of instances per modality. The summary of the number of instances per speaker, per modality is presented in Table 5.3. The large standard deviations indicate a large variation in the count of instances per subject – while some subjects may have very few instances others have several hundreds. All subjects have at least 5 instances in each modality.

5.5 Results

5.5.1 Feature distribution

Before analysing the classification results, we investigated some of the features, and compared them across controls *vs.* OSA patients.

Knowledge-based features for facial images Figure 5.6 shows statistics of the five KB features for the facial image modality, after zero mean and unit variance normalization, for OSA patients and the control group. By comparing the median values, we observe that the mandibular length, and the mandibular nasation angle are shorter in subjects with OSA than in controls, which is consistent with the findings of [Lee et al., 2009a]. On the other hand, for the three remaining features, we did not achieve results

consistent with those of [Lee et al., 2009a]: Lee et al. [2009a] reports, for subjects with OSA, a larger binocular width and a larger cranial base area, while we observe the contrary; [Lee et al., 2009a] also reports a larger face width for subjects with OSA, while we observe that OSA and control groups have approximately the same values.

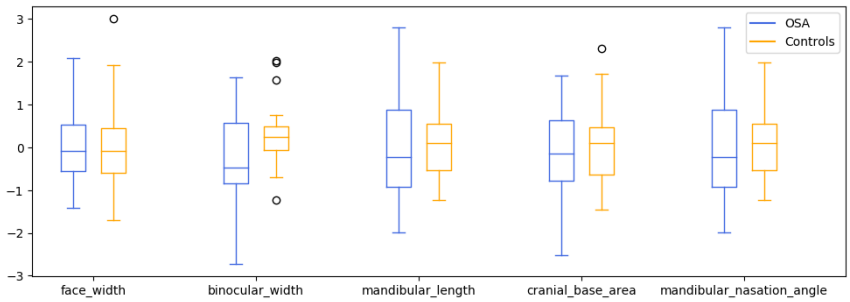
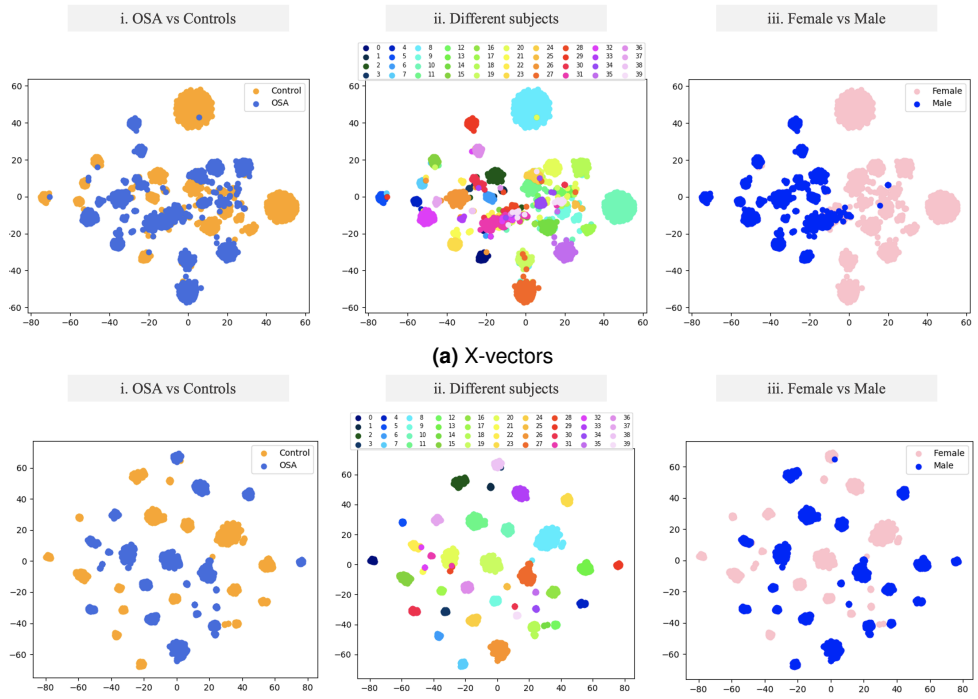


Figure 5.6: Boxplots of the normalized KB features, for OSA and control subjects.

The fact that our observations do not match the ones of Lee et al. [2009a] sustains the hypothesis that the KB features are not robust for OSA detection using facial images collected in-the-wild. The automatic computation of these features appears to be very sensitive to non-standardized poses of the face, occlusions induced by hand movements, beards or sun glasses, and the inconstant eye aperture in different frames of the videos which is used for calibrating the cranial base area. This variability introduced by in-the-wild data adds to the fact that the automatic detection of landmarks already introduces a relative error of 10% in the determination of the features, observed by [Balaei et al., 2018]. Furthermore, the fact that labels in our dataset may be noisy because subjects only claim to suffer from OSA and controls are chosen randomly, could contribute to the differences in our observations. This is valid not only for the KB features, but for all features.

Embeddings for identity recognition We plotted x-vectors and the embeddings designed for face recognition in Figure 5.7, using the t-SNE technique for visualizing high-dimensional data [Maaten and Hinton, 2008]. In the center plots, we mark each subject with a different color, and we observe that data naturally clusters around subject id. This is more evident for facial images than for x-vectors, which suggests that the audio data can be more noisy than the facial images - including background noise, music, etc., that may blur the subject clusters. On the left plots, we observe that there is no distinct separation between OSA and control groups, for neither of the modalities. The figure also shows a comparison of male and female embeddings, on the right plots. We observe that, in this projection of the high-dimensional space, x-vectors seem to better capture the gender identity than face embeddings.



(a) X-vectors

(b) Face recognition embeddings

Figure 5.7: t-SNE [Maaten and Hinton, 2008] visualization of face embeddings and x-vectors, for OSA *vs* control subjects (left), different subjects (center), and female *vs* male subjects (right). Note that the numbers assigned to each of the subjects are just denoting different subjects, but are not IDs – subject 1 in a) does not necessarily match subject 1 in b).

5.5.2 Classification results

Table 5.4 presents the classification results for the experiments described in section 5.4. Considering that the training mode was leave-one-subject-out cross validation setting, we define three accuracy metrics: *accuracy* is computed with the ratio of correctly classified instances on all cross-validation folds, over all instances of the dataset; *mean accuracy per subject* results from computing the mean of the accuracies obtained in each fold (i.e. each subject); and *majority vote accuracy* results of first performing a majority vote over the predictions for all instances belonging to each subject, obtaining one single prediction per subject, and then computing the accuracy. This last metric leverages the fact that we have multiple instances for each subject, and compensates for the fact that some instances may be noisy or even contain different subjects.

Table 5.4 also shows the chance level, or *prior*, for each modality, which corresponds to having a classifier that always predicts the most frequent class in the dataset. The prior for accuracy differs across modalities because the number of instances also differs; the prior for subject-level accuracy metrics remains constant because we are assuming a classifier based solely on the number of subjects, which remains constant for all modalities.

When analysing Table 5.4, overall, we observe that the facial images modality achieves the best

Table 5.4: Accuracy results [%] of the classification experiments. The prior corresponds to always predicting the label of the class with the highest *a priori* probability in the training fold.

Modality	Experiment	Acc	Mean acc per subject	Majority vote acc
Speech	prior	50.8	55.0	55.0
	x-vectors	65.1	62.9	67.5
	PASE+	60.3	62.3	62.5
	KB	55.6	55.5	55.0
Facial images	prior	53.5	55.0	55.0
	images	76.1	70.6	75.0
	images (local-att)	**76.3**	**73.4**	77.5
	images (global-att)	58.1	60.9	65.0
	KB	55.7	51.9	57.5
	BIF	61.1	58.4	55.0
	face embeddings	68.9	63.6	65.0
Visual speech	prior	51.7	55.0	55.0
	LRE	69.8	69.6	**80.0**
Fusion	prior	55.0	55.0	55.0
	early fusion (NN)	67.6 [67.1, 68.1]	67.6 [67.1, 68.1]	71.7 [70.9, 72.6]
	late fusion	–	–	**82.5**

results at the instance level, and visual speech achieves the best results after performing a majority vote. The speech modality performs worse than the other two. This is consistent with the findings of Lee et al. [2009a], which revealed a weaker association of OSA with speech, than with craniofacial characteristics. We hypothesize that the contamination of several audio segments with music and noise may further contribute to the worse performance of the speech modality. The use of a VAD module designed for broadcast news may not be robust enough for the vlog domain and must be replaced in future work. Future work can also consider concatenating the audio segments, such that each audio segment is at least one minute long before extracting the audio features. Comparing the three speech representations, we observe that x-vectors outperform the other two representations. The neural models trained for extracting both x-vectors and PASE+ features were trained with data augmentation, and thus may be more robust in a in-the-wild setting when compared to the KB features.

Regarding the facial images modality, we also observe that the KB features achieve the worst results at the instance level, in this in-the-wild setting. These results were expected, given the variability introduced by in-the-wild data collection previously described. BIF achieve better results than KB features at image level, but worst results after majority vote. As mentioned in section 5.3.3, BIF were designed to capture subtle variations in the texture of the skin, which although may have the potential to reveal OSA-induced sleep deprivation, are likely hidden by the YouTube content creators, using makeup and/or video editing. Face embeddings achieve better results than KB features and BIF, most likely because they capture more macro craniofacial morphology, which plays an important role in OSA detection.

Moreover, they might be more robust to manual manipulation, such as makeup. Feeding directly the raw images to a convolutional neural network outperforms all other forms of representations. When using the local attention layer, the accuracy reaches 76% at the instance level and 78% after majority voting. The results obtained with the local attention are better than the accuracy results reported by [Balaei et al., 2017] (69.8%) and [Nosrati et al., 2016] (73.3%), although results are not directly comparable, since datasets used in those works are different.

The embeddings trained for lip reading used to represent the visual speech modality outperform the other modalities, when comparing the majority vote accuracy.

The late fusion of the modalities by simple majority voting of the best three systems is able to slightly improve the performance, when compared to single modalities. The same does not happen for the early fusion experiment. However, the fact that we were able to provide 95% confidence intervals increases the trust in the early fusion results.

5.5.3 Interpreting predictions

Figure 5.8 shows the majority vote predictions attributed to each subject, for each modality. Overall, the three modalities tend to guess correctly more often the health status of OSA patients than controls. We also see that different modalities fail at classifying different subjects.

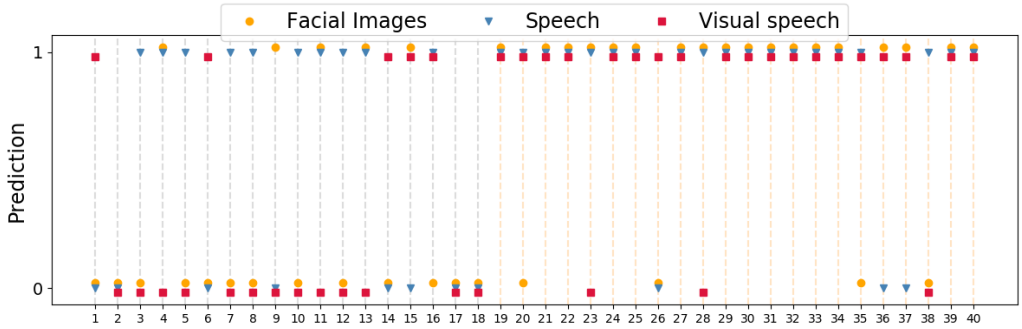


Figure 5.8: Rounded predictions per subject, using each modality. Subjects 1 to 18 are the controls, and 19 to 40 are OSA patients.

The best performance at the instance level was achieved using facial images, with the system described in *experiment B*. Thus, in this section, we analyse that system's predictions, in terms of misclassifications and investigating the attention weights.

Who are the false positives and false negatives? considering that OSA has a much higher prevalence in overweight subjects, it was important to make sure that the facial images-based classifiers were not simply detecting overweight subjects, and classifying them as OSA patients. In Figure 5.9, we show who are the false positives and the false negatives. We observe that the shown false positives are not overweight subjects, implying that our model is not an overweight detector.

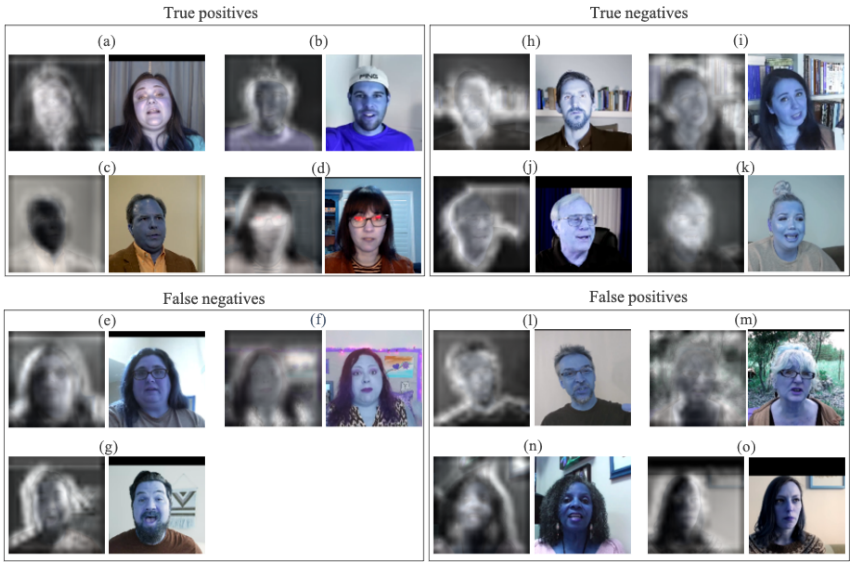


Figure 5.9: Examples of the attention scores learned by the system in experiment B of the facial images modality. The brighter pixels represent regions interpreted as more relevant by the system.

What does the model attend to? In an attempt to shed light on the model's decision process, we plotted the attention weights on top of each image, to understand which regions were considered more relevant. Figure 5.9 shows some of these plots, side-by-side with the input image. The input image has the colours altered due to a PyTorch operation, but the colour information is not relevant for this analysis.

To compute these plots, first we computed the mean of the attention weights across the channels dimension, then we upsampled the attention weights to match the height and width of the input image, using a bicubic interpolation. Afterwards, we added the upsampled attention weights to a black image with the same dimensions as the input image, to obtain the attention filter. For facilitating visualization, we normalized the attention filter between 0 and 1. Finally, we overlayed the normalized attention filter, multiplied by an α, on top on the original input image, multiplied by a β. We used $\alpha = 0.85$ and $\beta = 1 - \alpha$.

On the figures with the attention weights – left-side images on Figure 5.9, the brighter pixels represent areas considered more relevant for OSA detection. For most images, the higher scores were given to the area of the face and neck and/or edges around the face and neck. However, there are some unexpected cases – e.g. (c) and (i) – where higher scores were given to the background, which should not be relevant for the final decision. Preliminary experiments with of-the-shelf background removal methods did not yield better results for the classification task.

5.6 Summary

In this chapter, we explored three different modalities – acoustic speech, facial images and visual speech – for detecting obstructive sleep apnea, using in-the-wild vlogs. Our findings highlight the potential of

visual speech as a complementary modality for OSA detection. We anticipate that this preliminary investigation could extend to other speech-affecting disorders, particularly Parkinson's disease and depression. These conditions are marked by reduced emotional expression and psychomotor retardation affecting articulation. Visual speech embeddings have the capacity to capture information on articulatory and breathing patterns, as well as facial expressions of emotion, potentially aiding in the detection of these diseases. Notably, within our research group, a Master's Thesis has explored visual speech in the context of Parkinson's Disease [Ferro, 2023]. While, to the best of our knowledge, no studies have specifically investigated visual speech for depression detection, numerous works have employed multimodal approaches integrating speech and video (e.g. [Nasir et al., 2016]).

In this work, we compared KB features and transfer learning strategies. While past studies have commonly used KB features, in our in-the-wild setting, these features did not yield good results with speech or facial images. Improving the pre-processing and data selection might enhance their effectiveness but could also reduce the already limited amount of data. For instance, incorporating a face recognition system to ensure selected frames belong to the main subject should be considered, as the current outlier removal step addresses this only partially. Additionally, augmenting the facial images KB feature set with measurements like neck perimeter or double chin identification could be beneficial.

Transfer learning approaches, namely using embeddings for identity recognition, outperform the KB features. However, given our small sample size of 40 subjects, these embeddings may also overfit the training subjects and struggle to generalize properly to unseen subjects. The results might be improved by fine-tuning the networks that extracts the embeddings for the OSA detection task, instead of freezing the networks and using the embeddings as features.

Further improvements could be achieved with more extensive training and independent testing data. Data collection could be enhanced to include diverse age ranges and racial representations in both OSA and control groups.

Notwithstanding the good results obtained with end-to-end models, it is important to note that some of the attention plots show higher scores for background regions, which raises some questions on what are these models really learning. We cannot be absolutely sure that this system is not yet another example of anthropomorphism or shortcut learning [Geirhos et al., 2020], but we can emphasize the fact that, for most examples, the attentions scores focus around the face and neck, as expected. Furthermore, the results are in general very promising, compared to previous research, and considering the extra complexity introduced by the fact that data was not collected in a standardized approach, but rather in-the-wild, with strong variability.

Another avenue worth exploring is comparing the results obtained with the facial images modality with the results that could be achieved using educated guesses based on the intuition of medical experts. It will be interesting to investigate whether the machine learning systems make mistakes comparable to

those of medical experts.

Additionally, we conducted experiments to discuss the validity of using YouTube vlogs with other two diseases, depression and Parkinson's disease, which are detailed in Appendix B. The results achieved with cross domain experiments support the idea that the self-reported health status is a good proxy for the true health status, and also that this type of data allows us to model the diseases and not merely emotional or paralinguistic content that may be evoked when subjects are talking about their own disease. However, given the high prevalence of undiagnosed OSA, as discussed in chapter 2, caution is advised when interpreting these findings. A final idea for a promising future work direction is to address the uncertainty inherent to in-the-wild labels by employing soft labels instead of one-hot encodings. Instead of assigning zero probability of disease to a supposed control subject, soft labels can effectively capture disease prevalence within specific sub-populations based on demographics. It is recognized that training with fixed labels amidst noisy annotations typically results in poorer generalization [Vyas et al., 2020], and adopting soft labels could mitigate this issue.

6

Disease Detection Across Datasets: Uncovering Hidden Challenges in Data

Contents

THIS chapter presents a collection of experiments that highlight some of the challenges associated with the datasets used by the speech community for disease detection. In particular, we discuss unexpected biases; the difficulty of translating models and results to new domains, including different recording conditions, speech tasks, and languages; and finally, we explore how much information about the recording conditions is encoded in the features typically used to detect diseases from speech. The work presented in this chapter has been peer-reviewed and published at international conferences. Particularly, the experiments in section 6.2 were published in [Solera-Ureña et al., 2021] (second author of the publication), those in section 6.3 were published in [Ablimit et al., 2022], and the experiments in section 6.4 were published in [Botelho et al., 2022].

6.1 Introduction

Some of the results we discussed in the previous chapter for OSA detection are very promising. These findings align with a growing body of research demonstrating substantial performance in the automatic detection of various diseases using speech signals. For example, Han et al. [2022] and Al-Hameed et al. [2016] reported accuracies of 94.2% and 94.7%, respectively, in the binary classification of Alzheimer's disease patients and controls using the DementiaBank dataset. Additionaly, Orozco-Arroyave et al. [2016] achieved up to 99% accuracy on the automatic classification of Parkinson's disease patients and controls. These references provide a few examples among many existing studies.

Notwithstanding the promising results reported, and the large potential of the machine learning systems allied with noninvasive biomarkers, the reviewed literature does not indicate any transition of these systems to commercial products. Arguably, the most important issue to address before such transition takes place is to ensure the generalisability and reliability of the obtained results. Often the datasets used to train such systems have three strong limitations. The first limitation is that datasets are small when compared to the size of the datasets used for other speech tasks, and may not be representative of an entire population. The second limitation is that they focus on a single disease versus healthy controls, whereas in real-life, and namely in the context of an aging population, the coexistence of multiple diseases in the same patient, or *multimorbidity* (discussed in section 2.3), tends to be the norm and not the exception [World Health Organization, 2016]. The third limitation is that often datasets are heavily influenced by the conditions in which they were recorded or collected, the type of speech task they present, and the language that is spoken. This means that it is difficult to translate results to new datasets, and to understand whether a performance drop observed upon out-of-domain validation is due to the changing conditions, or because the model was learning from spurious correlations in the first place, and not from characteristics indeed attributable to the disease.

This chapter illustrates these limitations with concrete experiments and analysis. It presents a col-

lection of works that focus on identifying problems and raising questions. The first work described in this chapter focuses on the automatic detection of COVID-19, in this case from cough signals. The COVID-19 pandemic deeply impacted the world we live in, and we observed that the scientific community worked together and contributed in the best possible way[1]. However, despite the many efforts and collaborations, the research field of using AI for disease detection failed to deliver a suitable screening test that could be deployed easily with limited costs. We exemplify one problem that we identified in the COVID-19 cough data, that questions the reliability and generalizability of the obtained results.

The second work consists of a qualitative and quantitative analysis of speech and language features derived from two different corpora with the aim of predicting early signs of dementia. The two corpora differ not only in terms of speech task, but also in terms of language and time frame: one is a longitudinal German corpus, and the other is a cross-sectional English corpus. The results achieved are also different, which raises the question: *what are the learning systems really learning?* Shouldn't the properties of speech that are attributable to a disease hold across different datasets?

Following on this last question, we interrogate just how different can datasets be, even when maintaining the same language and similar speech tasks? Or even, for the case of a longitudinal dataset, how different can the audio samples that were collected at different measurement times be? Thus, the third work described in this chapter focuses on distinguishing audio samples of healthy speech that belong to different datasets, using the speech features typically adopted in works that use speech for disease detection. Furthermore, we suggest that one possibility to tackle the first two above mentioned limitations of datasets (reduced size and multimorbidity) is cross-corpora studies. In one hand, cross-corpora studies would allow to perform out-of-domain evaluation of the models trained in each small dataset. On the other hand they could enable the simultaneous detection of multiple diseases. In this context, understanding whether it is possible to combine different datasets becomes a very relevant question.

6.2 Biased datasets: the case of COVID-19 detection

As previously mentioned in section 2.2.2, since 2019, society was deeply impacted by the COVID-19 world pandemic. In particular, the years 2020 and 2021 were marked by an increasing interest in developing reliable, cost-effective, immediate and easy to use tools that could help healthcare operators, institutions, companies, etc. to optimize their screening campaigns for COVID-19. Several researchers and institutions have made efforts to collect speech, cough and breath data from healthy controls and people affected by COVID-19. Included in these efforts is the Cambridge COVID-19 Sound database [Brown et al., 2020; Han et al., 2021], which was partially used for the ComParE 2021 COVID-19 Cough Challenge [Schuller et al., 2021]. These initiatives provided researchers from all over the world a valu-

[1] https://www.nytimes.com/2020/04/01/world/europe/coronavirus-science-research-cooperation.html, https://www.nature.com/articles/d41586-020-00905-9

able space to share ideas and findings and to compare their results in a common test-bed.

This section describes our contributions to the ComParE 2021 COVID-19 Cough Sub-challenge, where we leverage transfer learning to develop a set of COVID-19 classification subsystems based on deep cough representation extractors (TDNN-F and CNN embeddings, and PASE+ features)[2]. Most importantly, in these sections, we also discuss the challenges of the dataset, the unexpected spurious variables we could find, and how these challenges may compromise the results in the literature. More than criticizing the dataset, this section aims at raising awareness for data quality, and the importance of working towards transparent classification systems in the context of healthcare, that could inform against possible biases in the dataset.

6.2.1 Related work

During the pandemic, research on automatic detection of COVID-19 from speech or respiratory sounds builds on previous studies investigating similar methods for other respiratory diseases such as pertussis, asthma, pneumonia, and tuberculosis [Pramono et al., 2016]. Although there are no conclusive evidences, preliminary findings suggest the presence of specific acoustic signatures of COVID-19 in coughs and speech, potentially enabling detection even in asymptomatic individuals and differentiation from other respiratory illnesses. Due to the limited availability of COVID-19-labeled data, many approaches employ transfer learning, data augmentation, and class balancing techniques.

Previous work predominantly employs CNNs in various configurations. For example, Brown et al. [2020] propose a pre-trained VGGish model [Hershey et al., 2017] as a generic audio feature extractor. Bagad et al. [2020] and Imran et al. [2020] initially train CNNs for cough detection and subsequently fine-tune the networks for COVID-19 detection. Chaudhari et al. [2021] propose an ensemble consisting of two Deep Neural Network experts and a CNN, trained from scratch, for direct COVID-19 detection.

Besides CNN-based approaches, other models have also been used for COVID-19 detection. Pinkas et al. [2020] present a three-stage system comprising a self-supervised transformer that generates embeddings from the inputs, a set of Recurrent Neural Network classifiers specialized in one or multiple types of input, and a final SVM meta-model.

Some works also incorporate self-reported symptoms information. For instance, Han et al. [2021] encode these symptoms as one-hot vectors, which are combined with traditional speech features at the feature or decision levels. Additionally, Pal and Sankarasubbu [2021] describe a system involving a DNN that produces cough embeddings from traditional handcrafted features, and a transformer-based self-attention network that generates embeddings from symptoms and demographic data. These embeddings are then concatenated and passed through a fully connected layer to yield the final decision.

[2]The entire work published in [Solera-Ureña et al., 2021] on COVID-19 is presented here, for context and coherence, however it is necessary to highlight that my main contributions concern the experiments that use TDNN-F embeddings.

6.2.2 Corpora

This work uses two datasets, described in detail in section 3.3: the COVID-19 COUGH (C19C) corpus, provided for the ComParE 2021 COVID-19 Cough Sub-Challenge; and the COUGHVID corpus, that was used to train and/or fine-tune the transfer learning-based cough representation extractors. For both datasets, silence segments were removed using a modified version of a cough segmenter developed by the COUGHVID team[3].

In our preliminary analysis of the C19C corpus, we noticed that some files present a reduced bandwidth of 4 kHz, hypothetically corresponding to audio samples originally recorded at a sampling rate of 8 kHz. This condition certainly reflects the reality of many real-world applications. However, we noticed that all the narrow-band recordings in the train and development subsets correspond to the COVID-19 positive class. From our analysis of the baselines and our own systems, we consider that this might be affecting their performance by making the training process pay attention to this spurious condition. For this reason, we decided to create a second version of the dataset by removing all the narrow-band recordings in the original train and development subsets, even at the cost of reducing the number of COVID-19 positive examples. The resulting dataset, denoted as "$C19C_{fullband}$", contains 273 samples in the train subset and 223 in the development subset. The test subset is kept unchanged to keep the original definition and evaluation conditions of the challenge. Table 6.1 summarizes the number of samples in the original dataset, and in "$C19C_{fullband}$" version.

Table 6.1: Number of cough samples in C19C corpus, before and after excluding narrow-band files. *C19* refers to COVID-19 patients, and *HC* refers to healthy controls.

	All samples					**Narrow-band samples**				
	Train		Dev		Test	Train		Dev		Test
	C19	HC	C19	HC		C19	HC	C19	HC	
$C19C_{original}$	71	215	48	183	208	**13**	0	**8**	0	**8**
$C19C_{fullband}$	58	215	40	183	-	0	0	0	0	-

6.2.3 Method

In this study, we employ transfer learning from a set of deep cough feature extractors that generate TDNN-F embeddings, CNN embeddings and PASE+ features. Subsequently, we use three SVMs to distinguish individuals as either healthy controls or COVID-19 patients, based on these cough-derived features.

TDNN-F embeddings

The success of *x-vector* representations for the detection of different diseases from speech motivated us to explore their applicability to coughs, hypothesizing that *x-vector*-like embeddings trained using

[3] https://c4science.ch/diffusion/10770

coughs, instead of speech, are capable of encoding relevant information about the cough signal and transferring medically meaningful information. Based on this hypothesis, we implemented a reduced version of the TDNN-F based network [Povey et al., 2018], proposed for speaker recognition by [Villalba et al., 2020]. The network architecture and training parameters are detailed in Appendix D.

We used the COUGHVID dataset to train the TDNN-F embedding network, addressing the limited availability of COVID-19 labeled data for training from scratch by separating training in two steps:

1. **Initial training:** The network was first trained for age estimation (regression) and gender detection (binary classification) using COUGHVID recordings with known age and gender information. Data were divided into training (7145 recordings), development (1531 recordings), and test (1531 recordings) subsets. We hypothesized that, similarly to speaker identity, a representation relevant for age and gender estimation would also encapsulate health-related information.

2. **Fine-tuning:** The pre-trained network was then fine-tuned using expert pulmonologists' annotations in a multi-task classification framework involving five tasks: *cough type* (dry/wet), *presence of dyspnea* (yes/no), *presence of wheezing* (yes/no), *diagnosis* (healthy cough/lower tract infection/upper tract infection/obstructive disease/COVID-19), and *severity* (healthy cough/mild/severe). This step aligns more closely with COVID-19 classification than the initial age and gender tasks, and may provide more informative representations. These annotations were provided for a small fraction of the dataset, by one to four experts and consolidated via majority voting, excluding recordings with tied votes. This resulted in 1285 recordings for training, 265 for development, and 256 for testing.

Input features were 30 MFCCs computed every 10 ms from 25 ms-length frames. Cepstral mean and variance normalization was applied using a 3 s-length sliding window. These steps were performed following the *egs/voxceleb/v2* Kaldi recipe [Povey et al., 2011]. The resulting cough TDNN-F embeddings are 128-dimensional vectors.

CNN embeddings

We leverage transfer learning from the mentioned VGGish model by Hershey et al. [2017] trained on a vast corpus for audio classification. We use this model in two settings:

1. **Pre-trained network:** We extract the embeddings from the pre-trained network.

2. **Fine-tuning:** Prior to embedding extraction, we fine-tune the model for COVID-19 detection using a balanced subset of the COUGHVID dataset (680 positive and 680 negative cough recordings; 80% of this subset is used for training and 20% for development).

Inputs to the VGGish network consisted of log Mel-spectrogram features computed every 0.24 s

from 0.96 s-length segments. Further details on the network architectire, and training parameters are provided in D. The resulting cough CNN embeddings are 256-dimensional vectors.

PASE+ features

In our experiments, we used two PASE+ feature extractors, as detailed in section 3.2. The first PASE+ extractor was trained on Librispeech [Panayotov et al., 2015]. The second was trained from scratch on COUGHVID data. The resulting embeddings are 256-dimensional feature vectors are extracted for each 10 ms frame.

COVID-19 classification

Given the limited amount of COVID-19 annotated data, our approach leverages transfer learning to obtain rich representations of cough, as described before. Subsequently, SVMs are used on top of each representation, to classify healthy controls and partietns suffering from COVID-19. Given the different nature of these representations, their respective pipelines are slightly different. File-wise TDNN-F embeddings are directly fed to the SVM. The CNN-based extractor generates a sequence of embeddings for each recording, computed from cough segments of 0.96 s with a shift of 0.24 s. Here, the sequence of embeddings is fed to the SVM and a final decision is taken by majority voting. PASE+ features are generated every 10 ms, with a receptive field of about 150 ms. In this case, an average feature vector computed across the whole sequence of features is fed to the SVM classifier. These three SVMs were trained on both the C19C and $\text{C19C}_{\text{fullband}}$ datasets. Different kernels (linear/RBF), data normalizations (zero mean and unit variance/[0,1] range) and class balancing methods (none/downsampling of the majority class/class weighting) were explored. The optimal configuration and hyperparameters were determined based on development results using a grid-search. Finally, system dependent scaling factors (and an offset) are estimated through linear logistic regression on the development subsets to combine the soft decisions of each individual system. The regression approximates log-likelihood ratios, thus, a theoretically determined decision threshold can be used for making hard decisions.

6.2.4 Results

Table 6.2 shows the performance in terms of UAR of the ComParE 2021 CCS baselines [Schuller et al., 2021] and our own system. All systems were trained separately on both the C19C and $\text{C19C}_{\text{fullband}}$ training subsets and evaluated on the corresponding *dev* and $\textit{dev}_{\textit{fullband}}$ subsets. Reported test results correspond to our best individual systems trained either on the C19C or on the $\text{C19C}_{\text{fullband}}$ datasets – this distinction is marked with $^{+}$.

Upon evaluating our proposed transfer learning-based cough representations, it is observed that the highest performance on the $development_{fullband}$ is achieved using the PASE+ features trained with COUGHVID data. These features attain a UAR of 66.8% on $development_{fullband}$ and 64.1% on the test

Table 6.2: Performance results (UAR [%]) on the C19C corpus. The $^+$ denotes whether the system used to present test results was trained on C19C or on $C19C_{fullband}$ datasets.

System		$dev_{original}$	$dev_{fullband}$	$test$
ComParE 2021 CCS Sub-challenge Baseline	openSMILE	61.4	53.0	65.5
	$openXBOW_{2000}$	64.7	56.5	72.9
	DeepSpectrum+SVM	63.3	57.3	64.1
	$auDEEP_{-60\ dB}$	67.6	57.3	67.6
	End2You	61.8	-	64.7
	Fusion of Best	-	-	73.9
TDNN-F Embeddings	Trained $COUGHVID_{Step1}$	68.8	63.6	-
	Fine-tuned $COUGHVID_{Step2}$	68.1	62.3	-
CNN Embeddings	Pre-trained YouTube	66.9	62.4	-
	Fine-tuned COUGHVID	71.2^+	65.6	62.3^+
PASE+ Features	Trained Librispeech	63.1	61.7	-
	Trained COUGHVID	67.4	**66.8**$^+$	64.1^+
Calibrated Fusion	Fusion of experts	**72.3**$^+$	66.1	69.3^+

set, showing a 5.1% improvement over the PASE+ extractor trained on the larger Librispeech dataset. A comparison of the results obtained by the PASE+ representations on the $dev_{original}$ and $dev_{fullband}$ suggests a higher robustness of these features to this spurious correlation.

The fusion of the best TDNN-F (trained on $COUGHVID_{Step1}$), CNN (fine-tuned on COUGHVID), and PASE+ (trained on COUGHVID) classifiers achieves the best performance, among our proposed systems, on the test set.

It is also important to notice that an analysis of the results obtained on the $C19C_{fullband}$ dataset, which excludes the narrow-bandwidth files, reveals that our proposed approaches outperform all baseline systems. In fact, baseline systems achieve a maximum UAR of 57%, which is barely above chance level (50%). Furthermore, the best performing baseline on the test set, $openXBOW_{2000}$, achieves only 56.5% accuracy on $dev_{fullband}$. These observations suggest that the baseline systems likely rely on spurious correlations, such as the bandwidth, to make predictions. Consequently, these systems exhibit poor generalization to new or out-of-domain data, and thus would be ineffective in real-world applications. Moreover, had we not identified the bias during our initial data inspection, there would have been no indication that the systems were learning patterns unrelated to COVID-19.

One possible strategy to overcome biased datasets, such as the C19C, is to perform comparisons across different corpora. For example, it would be unlikely that all COVID-19 related datasets would contain reduced bandwidth audios only for the case of subjects suffering from COVID-19. Verifying whether results hold in out-of-domain validation, or whether the conclusions observed in one dataset match the conclusions obtained in a different dataset gives a step towards more reliable and robust results. Furthermore, considering that corpora for health applications are often very limited in terms of size and population representation, it is very relevant to further study the generalizability of the results.

6.3 Transferability of results across datasets and languages: the case of Alzheimer's disease detection

This section proposes a comparative analysis for detection of Alzheimer's disease (AD) from speech, in two different corpus: the *Interdisciplinary Longitudinal Study on Adult Development and Aging* (ILSE) [Sattler et al., 2015], which consists of German biographic interviews, and the English cross-sectional ADReSS corpus, containing picture descriptions.

Speech and language biomarkers are strong indicators of dementia, and provide a low-cost and widespread alternative for the assessment of cognitive states. Several researchers have proposed methods to detect dementia and AD from speech and language features (e.g. [Jarrold et al., 2014; Pompili et al., 2020a; Tóth et al., 2015; Weiner et al., 2016]), making use of different speech corpora and achieving promising results. Contrarily, very few studies perform cross-corpora comparisons. Arguably, the uncertainty about the generalizability of results and methods is one of critical issues to solve before deploying speech and machine learning based solutions in clinical applications.

Thus, in this section we propose to analyse dementia detection in parallel, for the two corpora. We explore the audio and text modalities using 8 distinct sets of features/embeddings. We further discuss the longitudinal dimension of ILSE corpus, and the inherent challenges of changing recording conditions over time. In summary, this cross-corpus and cross-lingual study aims at answering the following key questions:

1. Are the distributions of features considered informative for the detection of AD similar in the two distinct corpora?
2. Do the best performing methods for dementia detection in the ILSE corpus work in ADReSS?
3. What aspects should be taken into special consideration when performing this analysis in a longitudinal corpus versus a cross-sectional corpus?

6.3.1 Related Work

The ADReSS 2020 challenge [Luz et al., 2020] released a benchmark speech dataset balanced in terms of age and gender, for two tasks: AD speech classification, and neuropsychological score regression. The main goal of this challenge was to address the lack of standardisation that currently affects the field, and introduce a dataset on which the different approaches could be systematically compared. The ADReSS contains two baselines for AD classification. The acoustic baseline uses ComParE features [Eyben et al., 2013] and LDA classifier, achieving 62.5% accuracy; the linguistic baseline uses a set of 34 language outcome measures (e.g., duration, mean length of utterances, type-token ratio, open-closed class word ratio, percentages of 9 parts of speech) and LDA classifier, and achieved an accuracy of 75%. As described in [Luz et al., 2021], the best performing models presented in the ADReSS 2020

challenge achieved an accuracy of 89.6% [Yuan et al., 2020] and 85.45% [Syed et al., 2020], using both text-based and acoustic features, and only text features, respectively.

6.3.2 Corpora

In the experiments described in this section we compare the $\text{ILSE}_{\text{m145}}$ and ADReSS corpora, previously described in 3.3. Although $\text{ILSE}_{\text{m145}}$ includes interviews associated with three diagnoses: controls, AD and age-associated cognitive decline (AACD), we only conduct the classification experiments on AD versus controls, to allow a fair comparison with the ADReSS corpus

The speech recordings in ADReSS corpus, made available for the 2020 challenge [Luz et al., 2020], were segmented using VAD and later normalised. The dataset made available contained both full enhanced audio, and normalised audio chunks. In our approach, we used the full enhanced audio and the corresponding transcriptions. We divided data into training / test sets with 108/48 subjects, respecting the partitions proposed in [Luz et al., 2020].

6.3.3 Method

Pre-processing

ILSE data: Given that the interviews have not been segmented at speaker turns, speaker diarization [Weiner et al., 2016] was applied to exclude speech from interviewers.

ADReSS data: For each recording, we discarded the interventions of the interviewer, and processed only the participants' parts. Only plain text was retained in the transcriptions, excluding non-speech events such as pauses and laughter. This approximates the ADReSS transcriptions to the transcriptions available for ILSE, and to the output that can be generated by an Automatic Speech Recognition (ASR) system.

Automatic Extraction of Features for Screening

We extracted speech and language indicators to capture speaker characteristics as well as the content and form of the message being transmitted. The 8 sets of features described below were extracted at the interview/complete picture description level.

Linguistic Inquire and Word Count (LIWC) [Tausczik and Pennebaker, 2010] is a text analysis tool that counts words in psychologically meaningful categories. It has been applied to mark individual differences in cognitive processing, namely for dementia screening [Weiner and Schultz, 2018]. Each word is mapped to one out of N categories ($N = 68$ for German and $N = 64$ for English), using a language-dependent LIWC-dictionary. Each LIWC feature reflects the percentage of occurrences of each category. LIWC categories may differ slightly across languages, but for each language they

were created to capture linguistic processes, psychological processes, personal concerns, and spoken categories.

Part-of-Speech (PoS) tags group words with similar grammatical properties into the same PoS tag, capturing the grammatical structure of the participants' speech. The transcripts are tagged using a language-dependent TreeTagger [Schmid, 1999, 2013], and each feature corresponds to the percentage of occurrences of each tag. The dimension of this feature set is 55 for ADReSS, 57 for ILSE.

Perplexity [Frankenberg et al., 2019] reflects how easy or hard it is to predict the words in a text. Perplexity scores have been shown to correlate with MMSE scores of subjects that develop AD or mild cognitive impairments 10–12 years later [Frankenberg et al., 2019]. For each interview, we developed n-gram language models by taking 80% of the segments of the interview as training set and evaluating the perplexity of the developed language models on the remaining 20% (n = 1, 2 and 3). The set of features related to the language models has dimension 15.

PoS Perplexity: We also apply perplexity to the PoS tags to explore the complexity of grammatical usage, using a 5-gram language model. This feature set has dimension 23 [Frankenberg et al., 2019].

VAD based features are a set of 11 clinically interpretable features based on silence/pauses metrics used in [Weiner et al., 2016]. This set includes silence count, silence count ratio, mean silence duration, silence duration variance, median silence duration, mean speech duration, speech duration variance, median speech duration, silence to speech ratio, mean silence count, and silence rate.

ComParE, extracted using openSMILE [Eyben et al., 2010].

i-vectors, extracted using Kaldi's [Povey et al., 2011]. The i-vector was extractor trained on the AMI corpus [Carletta et al., 2005].

ECAPA-TDNN embeddings, 192-dimensional vectors, were extracted using the model made available by SpeechBrain [Ravanelli et al., 2021], pre-trained on Voxceleb data[4].

The models used for extracting i-vectors and ECAPA-TDNN embeddings were not re-trained specially for ILSE using German data, for two main reasons: (1) to the best of our knowledge, there is no comparable German dataset regarding data size and number of speaker, and (2) Studies have successfully used x-vectors trained with VoxCeleb for disease detection in a different language [Jeancolas et al., 2021; Moro-Velazquez et al., 2020]

Feature normalization

All feature sets except i-vectors and ECAPA-TDNN embeddings were normalized with zero-mean and unit-variance normalization. For $\text{ILSE}_{\text{m145}}$, the normalization was performed separately for the interviews that belong to the same measurement time. The rationale behind this choice is that the interviews

[4] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

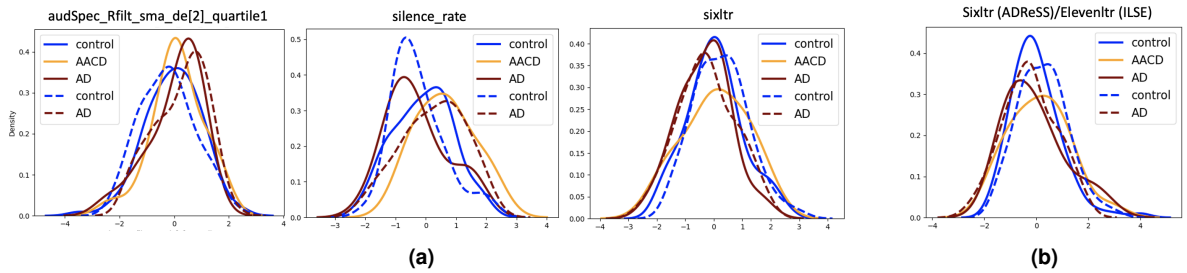


Figure 6.1: Density plots for feature distribution across both datasets. (a) Left: *audSpec_Rfilt_sma_de[2]_quartile1* from ComParE feature set; center: silence rate from VAD feature set; right: "six+ letter words" from LIWC feature set. The solid lines refer to ILSE and the dashed lines refer to ADReSS. (b) we plot the "six+ letter words" for ADReSS against a variation of that feature for ILSE, using eleven+ letter words.

at different measurement times are very distinct in terms of recording conditions, length (varies from roughly 1h to 5h per interview, excluding the interviewer segments) and discussed topics. Thus, this type of normalization was chosen to avoid the possible bias introduced by the fact that the majority of control interviews are from the first measurement period, and all the AD interviews are from the third one. I-vectors were scaled such that the L2-norm of each vector was 1.

Classification

We conducted binary classification to distinguish between patients with AD and controls, for each feature set using three classifiers: GMM with diagonal covariance, LDA and SVM with Radial Basis Function (RBF) kernel. Models were trained separately for each dataset using leave-one-subject-out cross-validation to cope with the limited dataset size. For the ADReSS data, performance was also evaluated on the held-out test set. Performance was measured using UAR. In the case of ADReSS, a balanced dataset, UAR is equivalent to accuracy, facilitating direct comparison with other studies reporting accuracy.

6.3.4 Results

Feature distribution

We qualitatively analyzed the distributions of the features for both datasets, using density plots. Figure 6.1 (a) shows only some examples of the feature distributions, for brevity. Regarding the feature *audSpec_Rfilt_sma_de[2]_quartile1* from ComParE feature set, we observe that the control distribution and the AD distribution are slightly shifted from each other, in the same direction for both datasets – which is a desired behaviour for features that could generalize AD detection across different corpora. On the other hand, the same does not happen for the features *silence rate* from VAD feature set, and "sixltr" (six or greater letter words) category in LIWC. For the silence rate, the distribution shifts seem to be inverted: ADs seem to have a higher silence rate when compared to controls in the ADReSS corpus, and a lower silence rate in ILSE corpus. The feature "sixltr" shows a distinct separation between AD

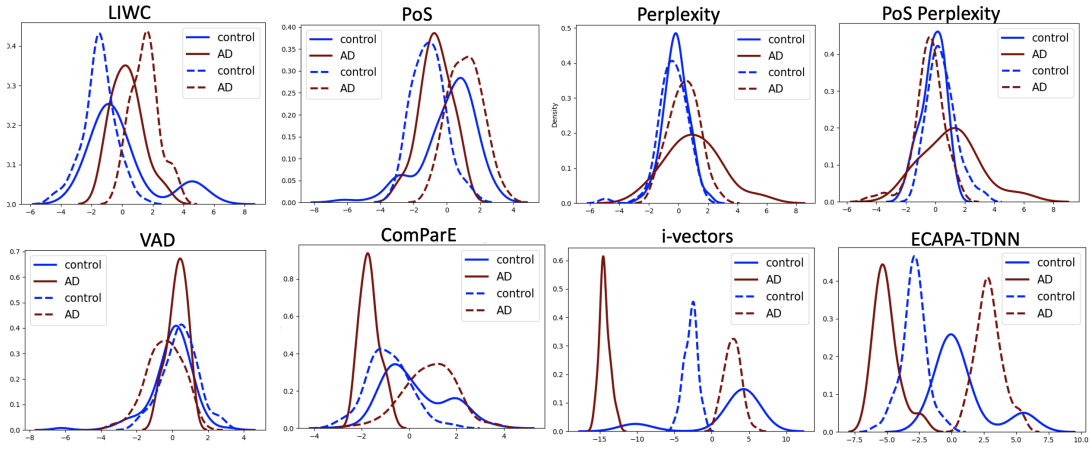


Figure 6.2: LDA projections of each feature set. The solid lines refer to ILSE and dashed lines refer to ADReSS.

and control distributions in the ADReSS dataset, but not in the ILSE dataset. This is likely because six-letter words are relatively long in English, often associated with formal speech, and it is understandable that people with cognitive deficits would use less of these words compared to healthy individuals. In contrast, German typically uses longer words, making "sixltr" a less effective indicator of cognitive impairment. To explore this hypothesis, we computed a new feature – "elevenltr" (words with eleven letters or more) – for the German ILSE dataset, and compared it with the "sixltr" feature for the English ADReSS (Figure 6.1 (b)). The rationale behind the choice of 11 letters is the fact that the average word length in English is 5.1 [Marian et al., 2012] and in German it is 10.6 [Admoni, 2019]. Comparing both plots, it seems that the 11-letter feature is better for German, than the 6-letter feature, because there is a smaller overlap between the control and AD distributions. Nevertheless, simply through a qualitative analysis, it is hard to say if the 11-letter feature for German is equivalent to what the 6-letter feature represents in English. These results call for future research to identify which features can be transferred across different languages and speech tasks, and how to adapt the features to make them meaningful across domains. Specifically, it may be beneficial to consider the number of phonemes rather than letters, as phonemes serve as more fundamental "building blocks" of words, which would entail the need tor phonetic transcriptions.

The large number of features does not allow an individual analysis of all of them, thus Figure 6.2 shows the one-dimensional LDA projection of each feature set. LDA projections are supervised (*control* vs. *AD*), and they were computed independently for each dataset, using all available data. Through visual inspection, one can observe that the projections of ECAPA-TDNN embeddings and i-vectors seem to have a smaller overlap between control and AD distributions, than the rest of the feature sets.

Besides the comparison of features across corpora, we also qualitatively explored how speaker embeddings vary across different interviews in $\text{ILSE}_{\text{m145}}$. For this end, we plotted in Figure 6.3 t-SNE projections [Maaten and Hinton, 2008] of ECAPA-TDNN embeddings extracted at the interview level for

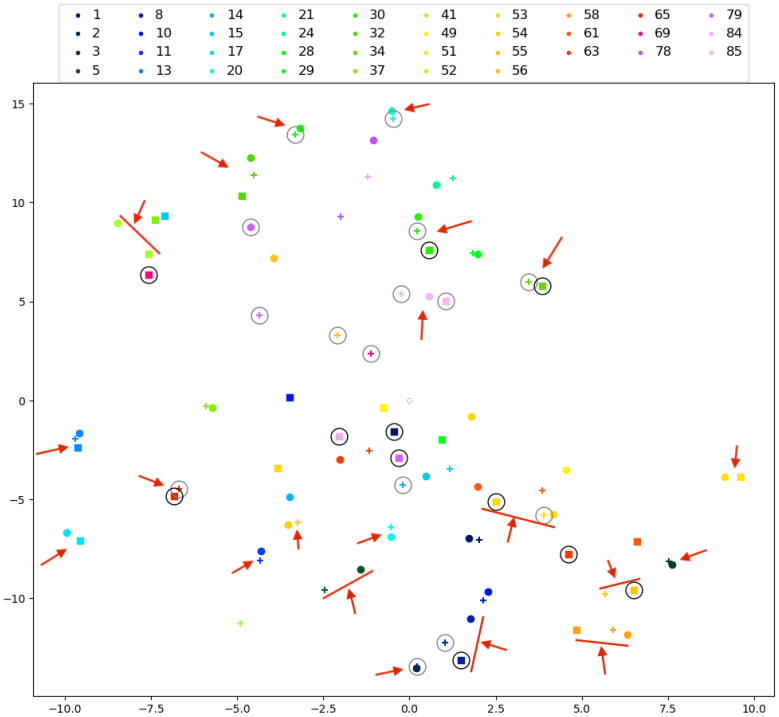


Figure 6.3: t-SNE representations of ECAPA-TDNN embeddings extracted at the interview level for all the 37 subjects in $ILSE_{m145}$ that have 2 or more interviews. Each subject is represented by a different color, and •, +, and ■ denote interviews that occurred at the first, second and third round of interviews, respectively. Black circles around the marks denote an interview where the subject has AD and gray circles denote an interview where the subject has AACD. Red arrows highlight subjects for which all their interviews were projected to the same region – showing a purely qualitative analysis.

the 37 subjects that have two or more interviews. Through visual inspection, we observe that the embeddings extracted from different interviews of the same subject sometimes (roughly for half of the subjects) lie in the same region of space, and sometimes are plotted far apart. This observation is not stronger in the cases where subjects develop AD or AACD, as it is possible to make the same observation (with roughly the same frequency) in healthy subjects. We thus hypothesize that the variability for the same speaker embeddings across different interviews is probably explained by larger margin through changes in the recording conditions and/or aging, which may obfuscate more subtle changes due to AD.

Classification

Table 6.3 summarizes the results obtained when performing the classification using each of the feature sets, for ADReSS and ILSE. For ADReSS, the acoustic features were able to achieve 66.7% of UAR, and the linguistic features achieve 77.1% of UAR on the held-out test set. These results surpass the baseline provided for the 2020 ADReSS challenge (62.5% and 75%, respectively) [Luz et al., 2020]. The linguistic features which are more informative are LIWC and PoS tags, while ECAPA-TDNN embeddings achieve the best performance when compared to other acoustic representations. SVM is the best classifier. Furthermore, we observe a large performance gap between training cross validation and the held-out

Table 6.3: Classification results in ADReSS and ILSE (UAR [%]).

	ADReSS						ILSE		
	GMM		LDA		SVM		GMM	LDA	SVM
	CV	test	CV	test	CV	test			
LIWC	69.4	60.4	99.1	68.8	99.1	**77.1**	78.9	39.2	48.1
PoS	64.8	54.2	89.8	68.8	92.6	**77.1**	**83.8**	41.2	38.0
Perplexity	55.6	54.2	68.5	45.8	73.1	56.2	55.9	55.8	71.6
PoS Perplexity	55.6	47.9	71.3	58.3	68.5	54.2	55.7	54.9	50.0
VAD	50.9	56.2	69.4	60.4	70.4	62.5	45.7	49.5	56.7
ComParE	94.4	47.9	75.0	60.4	99.1	58.3	49.5	69.0	34.0
i-vectors	55.6	50.0	60.2	65.5	56.5	54.2	84.6	79.3	**86.0**
ECAPA-TDNN	93.5	**66.7**	95.4	60.4	99.1	**66.7**	61.9	66.7	52.9

test set, which reflects some overfitting during training. Nevertheless, the models are still able to surpass baseline performance on the test set.

For the ILSE corpus, because we do not have defined a held-out test set, we report all results only in terms of leave-one-subject-out cross validation. The best results for the linguistic features (83.8% UAR) are achieved with the pair PoS tag features and GMM classifier, while the best result achieved with acoustic features is 86% with i-vectors and SVM classifier.

In general, when comparing the cross-validation results of ILSE with the cross-validation results of ADReSS, the former appear worst than the latter. Nevertheless, it is possible that the models are overfitting more the ADReSS data. Another important aspect to highlight is the fact that the best classifiers and the best feature sets for one dataset do not match the best for the other dataset, with the exception of the PoS tag features.

We also experimented with models trained in one dataset and evaluated on the other, i.e., we took a model trained on ILSE and tested on ADReSS and vice-versa. The results were barely above chance level. The exception occurred for the perplexity features, that achieved a UAR of 62.5% on the test set of ADReSS, using an SVM model trained in ILSE.

One should consider that the two datasets differ not only in language, but also in the number of subjects, total duration of the audio, and type of data. ILSE recordings correspond to German biographic interviews, while the ADReSS recordings correspond to the English description of the Cookie Theft. While it is true that both tasks trigger spontaneous speech, the area and diversity of vocabulary, as well as the emotional content are expected to differ across datasets. These facts may explain the different results found across datasets. Nevertheless, considering that datasets for medical diagnosis are typically small and heterogeneous, we argue that it is crucial that future research discusses how to translate results across different domains, and how to measure robustness and reliability of the results.

6.4 Healthy speech across corpora and time

In the previous section, although we achieved promising results for each dataset individually, the optimal feature sets differed between datasets, and models trained on one dataset performed poorly on the other. While the differences in terms of language, speech task, and recording conditions may explain the different results, we argue that it is crucial to discuss how to translate results across different domains, and how to measure robustness and trust of the results achieved in the research area of speech for health.

In this section, we discuss whether it is possible to combine different datasets for disease detection. This could be useful, for instance, to perform out of domain validation, i.e., to train the classifiers in one domain and test them in a different domain, to promote generalizability, reliability and robustness of results. Alternatively, it could also be useful to combine different datasets in the context of simultaneous disease detection, given that most datasets pertain a single disease. Finally, this discussion is also relevant in the context of long duration longitudinal studies, where it is only natural that recording conditions change over time.

Although ideally, acoustic-based systems for disease detection would be language-independent, recent studies have found that language-specific differences in aspects such as phonation and prosody may influence the perception of speech impairment [Despotovic et al., 2021]. For this reason, we restrain the problem to mixing corpora in the same language. Thus, we compare the speech of healthy subjects, speaking the same language, collected across different datasets, using features that are typically adopted in studies of pathological speech. Concretely, we design experiments to answer the following questions:

1. How easy is it for a machine learning classifier to distinguish between recordings of healthy people collected for different datasets (or measurement time, in the case of longitudinal studies), using acoustic features typically used for disease screening from speech?
2. Are embeddings obtained with neural networks trained with data augmentation more robust to dataset changes than knowledge based features?
3. Can unsupervised clusters of speech data, represented by the typical acoustic features used for health classification, be mapped to the source dataset (or measurement time, in the case of longitudinal studies)?

6.4.1 Method

As explained above, this study compares the speech of healthy subjects (i.e., subjects for which there is no reported presence of disease) across different recording conditions. We define two studies. The non-longitudinal study, study 1, involves six corpora in American English. The longitudinal study, study 2,

pertains always the same corpus, ILSE corpus in German, but involves changes of recording conditions over four recording times. The comparison of speech samples over changing recording conditions was made on the basis of feature sets that are typically used in studies of disease detection from the acoustic signal, namely: eGeMAPS, ComParE feature set, i-vectors and ECAPA-TDNN embeddings, all of which have been previously described in section 3.2. This section briefly describes the datasets and the set up of each of our experiments.

Corpora

In the selection of the cross-sectional English datasets for this study, we involved spontaneous speech datasets, read speech datasets and datasets including both. All datasets are roughly balanced in terms of gender. As a starting point, we selected three data subsets corresponding to the healthy control subjects of corpora collected for detecting diseases from speech: AD (ADReSS), PD (WSM), and Depression (DAIC-WOZ and WSM). To complement these subsets we took CLAC, a dataset collected on purpose to serve as a speech corpus of healthy English speakers and two other corpora totally unrelated to the study of diseases from speech: TIMIT and VoxCeleb. The criteria for choosing the last two corpora, TIMIT and VoxCeleb, was the availability of age information that could be explored in later studies.

All datasets used in both study 1 and study 2 have been previously described in section 3.3. ADReSS and ILSE were also used in the previous section.

ADReSS, includes speech recordings of 78 healthy control subjects. Our study was restricted to the subset of this corpus used in the ADReSS challenge. Although the version used in the challenge was acoustically enhanced, we opted for using the original version from the Pitt corpus [Becker et al., 1994] in order to avoid the tonal noise.

CLAC's speech tasks considered in this study correspond to the reading passages and picture descriptions.

DAIC-WOZ, includes speech recordings of 100 healthy subjects.

TIMIT, contains speech from 630 speakers, but one subject was excluded due to lack of age information. The recordings of the 10 reading sentences of each subjected were concatenated into a single file, to provide an average segment duration more comparable to the remaining datasets.

VoxCeleb [Nagrani et al., 2019] includes short clips taken from interviews uploaded to YouTube. Hechmi et al. [2021] have annotated a subset of VoxCeleb 2 with age, gender, nationality, among other information. In this study, we included only the interviews of subjects with available age information from the USA, to avoid including interviews that are not spoken in English. This results in a subset of 840 subjects.

WSM, includes 587 subjects which did not claim to suffer from PD or depression, and for whose vlog

Table 6.4: Corpora description.

	#Subjects			#Segments			Age		Duration	
	M	F	O	Train	Dev	Test	range	mean±std	mean[s]	total[h]
Six English datasets:										
ADReSS	35	43	0	46	8	24	50-78	66±7	58	1.3
CLAC	894	900	13	4351	939	930	18-80	36±12	35	61
DAIC-WOZ	61	39	0	525	85	66	NI	NI	72	14
TIMIT	438	191	0	392	69	168	20-75	30±8	31	5
VoxCeleb	477	363	0	2492	73	140	10-80	44±15	45	34
WSM	289	298	0	3384	545	842	18-70	37±10	79	105
ILSE:										
Time 0	52	37	0	10303	1628	1899	44-63	55±9	71	273
Time 1	47	42	0	5683	1003	1132	48-68	65±5	68	146
Time 2	45	44	0	3510	579	754	55-75	63±9	67	90
Time 3	45	44	0	2751	532	466	62-83	69±9	68	64

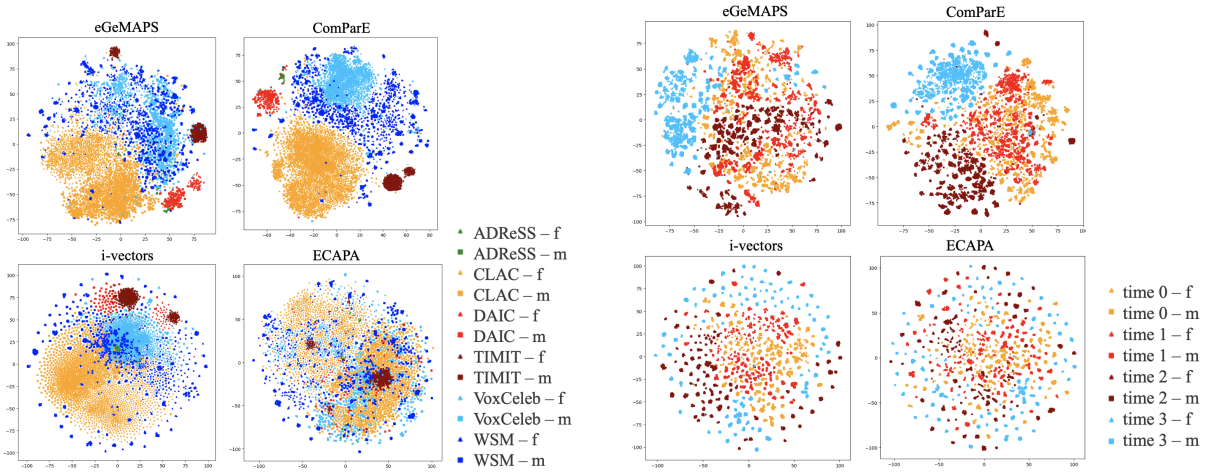
only one subject appears in the entire video. The audio recordings were segmented using an in-house VAD [Meinedo and Neto, 2005].

ILSE, is a German corpus of biographic interviews. Although each participant engaged in up to four measurements, in this study we used a single interview from each speaker, amounting to 356 speakers equally divided into 4 subsets. Naturally, the age of the subjects in the longitudinal corpus increases across the measurement times, but because the corpus includes two cohorts with distinct birth dates, there is always some age overlap between each of the measurement times.

We respected the train/dev/test set partitions proposed for these corpora, unless the partitions were not made available, in each case we randomly split the data to define the partitions. In ADReSS and DAIC-WOZ, we excluded speech from interviewers. In interview/vlog datasets and TIMIT, we concatenated several participants turns such that each audio segment would have roughly one minute of audio. Table 6.4 summarizes the number of subjects and segments, as well as age and duration information for each of the datasets used. We normalized the eGeMAPS, and ComParE feature sets with zero mean and unit variance, based on the train set, for all experiments except when a different normalization strategy is explicitly described.

Experiments

Experiment A: Supervised classification of healthy speakers from distinct datasets – We perform classification using SVMs and each of the feature sets described above. The SVM hyperparameters (kernel, *C*, gamma and degree) were chosen with a grid search on the development set. The goal of this experiment is to discuss how easy it is for a learning system to detect the dataset. In experiment *Experiment A-1*, we perform 6-class classification, using the 6 English datasets. In experiment *Experiment A-2*, we perform 4-class classification to distinguish between the 4 measurement times of the longitudinal corpus ILSE. Although we expect that this experiment is harder than experiment A-1, it is just as relevant



(a) Six English datasets

(b) Four ILSE Measurement times

Figure 6.4: t-SNE projections of each of the feature sets, for the 6 English datasets (a) and 4 measurment times of ILSE (b). ■ for male and ▲ for female.

because it is natural that recording conditions change over the course of large longitudinal corpus, and the participants of the studies may only develop diseases at later measurement times. Hence, recording conditions will be a strong confounder when trying to detect diseases from speech in such data.

Experiment B: Clustering – We perform *agglomerative clustering* to assess whether, even without supervision, data would naturally cluster around source dataset. We hypothesize that the gender of the subjects may also play a relevant role when forming the clusters, thus we compare three different numbers of clusters $\{2,$ *number of classes*, $2 \times$ *number of classes*$\}$, where *number of classes* corresponds to the number of datasets or measurement times. In *Experiment B-1*, we perform clustering over a pool of data that includes all 6 English datasets. In *Experiment B-2*, we perform clustering over ILSE data.

6.4.2 Results

Exploring the features Figure 6.4 (a) and (b) show the t-SNE projections of each of the feature sets, for the six English datasets and ILSE, respectively. We observe that eGeMAPS and ComParE feature sets allow the projection of the different datasets/measurement times to clearly separate regions of space. On the other hand, ECAPA-TDNN appears to provide a less clear separation between datasets. It is also visible that gender plays an important role. For example, TIMIT (in brown) forms two clear clusters, one for male and one for female speakers, in all feature sets.

Classification results Supervised classification results are reported in terms of accuracy, and unweighted average recall (UAR), Unweighted Average Precision (UAP), and Unweighted Average of F1 score (UF1) across the classes. We emphasise UAR analysis, given that it is the standard metric used

Table 6.5: Experiment A-1: classification of six distinct datasets using features often used for disease detection from speech.

	Dev [%]				Test [%]			
	Acc	UAP	UAR	UF1	Acc	UAP	UAR	UAF1
eGeMAPS	89.3	84.4	92.8	86.8	90.6	88.8	89.3	88.4
ComParE	99.8	99.6	97.9	98.6	98.6	98.7	94.8	96.6
ECAPA-TDNN	83.7	81.6	79.8	77.2	86.1	77.9	82.0	78.2
i-vectors	94.6	90.8	94.2	91.4	89.0	88.7	89.1	87.5
After dataset-dependent normalization:								
eGeMAPS	95.3	92.5	84.8	86.5	92.7	91.2	81.2	83.5
ComParE	96.7	96.3	79.4	81.8	90.9	91.6	73.4	74.4
ECAPA-TDNN	81.2	53.8	50.6	48.3	73.3	53.0	50.2	47.9
i-vectors	76.7	65.8	53.0	49.7	72.7	67.4	52.4	48.7

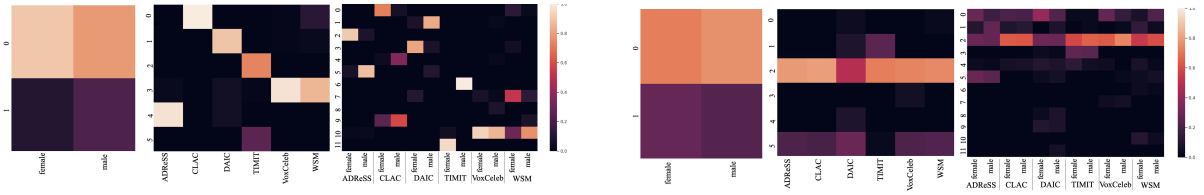
for disease classification from speech, because the vast majority of datasets have an imbalanced class ratio [Cummins et al., 2018]. The results of the classification experiments A-1, using the six English datasets are presented on the top part of Table 6.5. All UAR results are above 82%, and the ComParE feature set reaches 95% UAR. These results show that dataset classification is fairly easy, using these standard feature sets. We expected that ECAPA-TDNN embeddings would contain less information relevant for dataset classification, because they are discriminative representations trained with data augmentation, thus should be more robust to small domain shifts. The results confirm this expectation, i.e., ECAPA-TDNN embeddings achieve the lowest results over all metrics, but the performance gap is not as large as could be expected. i-vectors, on the other hand, model total variability, which can justify the fact that they contain more information specific to the dataset conditions, such as the speech task or recording conditions.

The bottom part of Table 6.5 shows the results after performing dataset dependent normalization of the features. We hypothesize that this normalization may remove some dataset-specific information, and make the feature sets more dataset agnostic. In fact, the performance drop after normalization confirms that hypothesis. This performance drop is particularly evident for ECAPA-TDNN embeddings and i-vectors. Nevertheless, this normalization does not remove all dataset-specific information, given that results for all feature sets are highly above chance level (16.6% UAR).

Table 6.6 presents the results for experiment A-2, which involves classifying the ILSE corpus into the four distinct measurement times, before and after measurement time-dependent feature normalization. Consistent with the findings from experiment A-1, the ECAPA-TDNN feature set performs worse than the remaining feature sets, when trying to classify measurement time. Regarding the effects of the measurement time feature normalization, for most feature sets (i-vectors, ECAPA-TDNN and ComParE) we observe a performance drop post-normalization. However, this performance drop is not as pronounced as in experiment A-1. Additionally, the reverse trend is observed for eGeMAPS. This observation is likely

Table 6.6: Experiment A-2: classification of the four measurement times in ILSE, using features often used for disease detection from speech.

	Dev [%]				Test [%]			
	Acc	UAP	UAR	UF1	Acc	UAP	UAR	UF1
eGeMAPS	89.7	89.9	89.7	89.7	85.1	86.0	85.1	85.3
ComParE	92.1	92.1	92.1	92.1	94.7	94.9	94.7	94.8
ECAPA-TDNN	67.1	67.8	67.1	67.1	67.7	69.0	67.7	68.0
i-vectors	90.3	90.8	90.3	90.5	88.6	89.9	88.6	88.8
After measurement time-dependent normalization:								
eGeMAPS	88.7	88.6	88.7	88.5	91.8	91.9	91.8	91.8
ComParE	93.8	93.8	93.8	93.7	94.2	94.5	94.2	94.2
ECAPA-TDNN	72.6	79.4	72.6	71.9	65.9	76.3	65.9	66.9
i-vectors	81.7	83.1	81.7	81.9	81.7	83.3	81.7	81.6



(a) No normalization

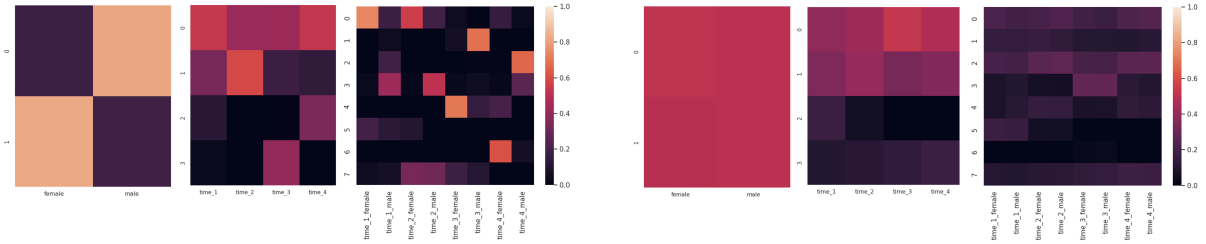
(b) Dataset dependent normalization

Figure 6.5: Agglomerative clustering of the 6 English datasets, using i-vectors. From left to right: 2, 6 and 12 clusters.

attributable to the fact that the different measurement times within the ILSE corpus are more similar to each other, than the six English datasets used experiment in A-1. Consequently, the impact of separate normalization is less pronounced in this context.

Clustering results Figures 6.5 and 6.6 show the results for the clustering experiments B-1 and B-2, respectively, using i-vectors as input features. We opted to display i-vectors because they seem to provide the most clear results. In each heat map displayed in figures 6.5 and 6.6, each row corresponds to one predicted cluster, and each column to a characteristic that we hypothesize that could be associated with each cluster. The colors correspond to the normalized frequency of observations of the given characteristic in each of the predicted clusters, thus the values in each column sum to 1. The most left heat map corresponds to clustering into two clusters; the center corresponds to the number of cluster equal to the number of classes (six datasets in Figures 6.5, and four measurement times in Figure 6.6); and the right heat map corresponds to $2\times$ number of classes.

Analyzing Figure 6.5 (a) we observe that when clustering into two clusters, there is no predicted cluster that corresponds mostly to female (nor male) speaker. On the other hand, when clustering into six clusters, we do observe that each cluster corresponds roughly to each of the datasets. The exception occurs for WSM and VoxCeleb, which are placed into the same cluster – this makes sense considering that both WSM and VoxCeleb are collected from YouTube, although the former corresponds



(a) No normalization **(b)** Dataset dependent normalization

Figure 6.6: Agglomerative clustering of ILSE data, using i-vectors. From left to right: 2, 4 and 8 clusters.

to vlog format and the latter to interview segments. TIMIT is split into two separate clusters, possibly one corresponding to male and the other to female subjects. In the the 12-cluster heat map, we observe that each cluster corresponds mostly to either male or female subjects of a single dataset, with the exception of WSM and Voxceleb.

After dataset dependent normalization of i-vectors, Figure 6.5 (b), we observe that each cluster cannot be mapped anymore to single dataset. These results are consistent with the performance drop for i-vectors observed in Table 6.5 after normalization.

In Figure 6.6 (a), we observe that when the number of clusters for ILSE is two, there is a tendency to cluster according to gender. The 8-cluster heat map shows some mapping between predicted cluster and datasets. We observe that several of the female speakers of times 1 and 2 are grouped into cluster 0, and the male speakers of the same times are grouped into cluster 3. Most of the remaining clusters correspond to to a single gender, and a single measurement time. After measurement time dependent normalization of i-vectors, Figure 6.6 (b), any existing mapping between clusters and measurement times or gender vanishes. Although in Table 6.6 there is also a performance drop after i-vector normalization, Figure 6.6, (b) might suggest even worse supervised classification when applying normalization, than the results actually obtained.

It is important to highlight that the results in Figures 6.5 and 6.6 are completely unsupervised, and the fact that, in figures (a), data naturally cluster according to dataset or measurement time shows that these feature sets encode much information about the recording conditions/task. Furthermore, the absence of defined clusters in figures (b) does not mean that dataset (or time measurement) specific information is completely removed by the normalization. It just means that it becomes less evident in the context of unsupervised learning.

Figure 6.7 shows the euclidean distances between the mean i-vector of each of the six English datasets, and of each of the four measurement times in ILSE (right). We observe that the smaller distances occur for the pairs WSM-VoxCeleb and measurement time 0 and 1, which is consistent to what we observed in the cluster figures. Generally, we also observe that the distances between measurement times in ILSE are smaller than distances between distinct datasets.

	ADReSS	CLAC	DAIC	TIMIT	VoxCeleb
ADReSS	0				
CLAC	0.52	0			
DAIC	0.44	0.47	0		
TIMIT	0.61	0.55	0.43	0	
VoxCeleb	0.42	0.41	0.38	0.48	0
WSM	0.45	0.33	0.41	0.51	0.22

	Time0	Time1	Time2
Time 0	0		
Time 1	0.20	0	
Time 2	0.31	0.33	0
Time 3	0.25	0.32	0.36

Figure 6.7: Euclidean distances between mean i-vectors of each of the six English datasets (left), and of each of the four measurement times in ILSE (right).

6.5 Summary

The experiments described in this chapter highlight some of the challenges of the current datasets used by the speech community for disease detection. In particular, we started by describing our experiments using the C19C corpus for COVID-19 detection, and we discuss how the corpus is strongly affected by a confounding condition - the bandwidth. This work exemplifies unexpected biases in datasets and how those biases can question the validity of black box machine learning-based classifiers. We also discuss our experiments that target AD automatic detection from speech in a longitudinal conversational corpus, spoken in German – the ILSE corpus – and establish a comparison with a publicly available cross-sectional corpus, spoken in English – the ADReSS corpus. Although we achieve promising results for each dataset individually, we observe that the best classifiers and the best feature sets for one dataset do not match the best for the other dataset, which raises questions on the transferability of the methods to new domains, such as different languages, different tasks and different recording conditions.

Lastly, we explore healthy speech collected in different datasets and different recording times of the same longitudinal dataset, using the same features typically employed for the detection of diseases from speech. We show that all feature sets analysed encode substantial information about the dataset/recording conditions over time. We support this claim through supervised learning experiments with results largely above chance level, and through unsupervised experiments where data naturally clusters according to the dataset/measurement time. These experiments emphasize the importance of understanding our classifiers, and making sure that disease detection is not performed based on specific dataset characteristics. They open the discussion on how can we combine different datasets, both for out-of-domain validation and for multi-disease classification.

The experiments described in this chapter call for future research to explore, on one hand, how can we be sure that what we are learning is indeed attributable to the disease and not to aging, differences in recording conditions or other characteristics of the dataset; and on the other hand, which methods can we use to leverage data from different domains/datasets to build trustworthy models that address health from a holistic perspective. One could argue that collecting larger datasets, in controlled conditions, covering different diseases and languages could solve or minimize several of the challenges discussed

here. However, considering the difficulty of collecting such a dataset, in terms of resources and ethical concerns, we believe there is space for improvement using interpretable and trustworthy solutions that leverage already existent data.

7

A Framework towards multidisease screening

Contents

THE previous chapter called for research in the direction of leveraging different datasets, each annotated for a single disease, to develop a robust health monitoring model. This model should effectively handle dataset shifts, consider multiple diseases, and provide explanations aligned with clinical reasoning. Hence, we propose that a valuable first step is to characterize the speech of control subjects, and to work towards a definition of reference speech that can be used independently of the dataset of origin.

This chapter presents a framework that first defines reference speech, and then leverages this definition to perform disease detection. Reference speech is characterized using reference intervals for clinically meaningful acoustic and linguistic features derived from a reference population. This novel approach in the field of speech health is inspired by the use of reference intervals in clinical laboratory science for medical diagnostics. We then quantify deviations of new speakers from this reference model (deviation-scores), and use these deviations as input to detect Alzheimer's and Parkinson's disease. One classification strategy explored is based on Neural Additive Models, a type of glass-box neural network.

The framework proposed in this chapter for reference speech characterization and disease detection is designed to support the medical community by providing clinically meaningful explanations that can serve as a valuable second opinion.

The experiments detailed in this chapter resulted in two manuscripts. One manuscript was published at Interspeech 2023 [Botelho et al., 2023], and the other was published at the IEEE Access Journal [Botelho et al., 2024].

7.1 Introduction

In the quest for a more trustworthy use of speech as a biomarker for disease screening, we aim to address two main aspects. The first aspect concerns the need to ensure that disease classification models learn properties indeed attributable to the disease, and not other confounding factors. In fact, this problem has also been identified by other works. For example, Berisha et al. [2022] discussed the association between small sample sizes and inflated results. They found that accuracy in classifying healthy controls and dementia patients from speech data increased as sample sizes decreased, suggesting publication bias and overfitting as the main contributing factors. In a recent survey talk[1], Cummins has also discussed this problem and cautioned against the Clever Hans Effect in this context. Ozbolt et al. [2022] identified several methodological issues that could lead to overoptimistic results in classifying PD patients and healthy controls using sustained vowels, including age-unmatched classes, large feature vectors, and data leakage between train and test sets. Espinoza-Cuadros et al. [2016] highlighted two

[1] N. Cummins. Machine Learning for Speech-based Health Analysis: State-of-the-art and Future Challenges. Survey talk at Interspeech 2022 https://www.interspeech2022.org/program/surveytalk.php

main methodological limitations in obstructive sleep apnea detection from speech analysis: the influence of confounding factors, such as age, height, sex, etc., and overfitting of feature selection and validation methods when working with a high dimensional feature set compared to the number of samples. The frequently reported overoptimistic results due to confounding factors and overfitting on small datasets call for more reliable research on the use of speech as a biomarker for disease screening. Namely, using interpretable models may represent a step towards ensuring that the model is learning properties indeed attributable to the disease, and not other confounding factors.

The second aspect that we aim to address relates to multimorbidity, and the fact that speech affecting diseases often have overlapping effects on the speech signal, and are frequently considered risk factors for each other, as previously discussed in chapter 2, section 2.3.

Based on these considerations, we hypothesize that a speech-based tool to support medical diagnosis and monitoring of chronic conditions, should adopt a holistic approach to health. This approach should facilitate an interpretative assessment across multiple diseases rather than relying on black-box models that offer a binary classification between a specific disease and healthy controls. However, existing datasets for disease detection are often limited in size and labeled for individual diseases. Consequently, as discussed in sections 6.3 and 6.4, the naive combination of different datasets containing individuals with a single specific disease to perform a cross-corpora study for multi-disease classification would most likely result in unreliable results that would not properly generalize to unseen recording conditions.

Conisdering this, we claim that a valuable step towards the adoption of speech and language technologies in real health applications would be to obtain a definition of reference speech that could be used independently of the dataset of origin, and later be applied to identify disease signatures. In this study, *reference speech* refers to the speech characteristics common to a reference population, ideally comprising healthy individuals of varying ages and biological sexes. Acknowledging the challenges of defining health and the prevalence of subclinical disease[2], we do not assert that our reference population consists exclusively of healthy speakers. Instead, we utilize the speech of individuals who self-report as disease-free. We propose to characterize reference speech through *Reference Intervals (RIs)* of clinically meaningful speech and language features. RIs represent the typical range of values for specific parameters within a reference population. In this context, RIs are computed as the 2.5th and 97.5th percentiles of the distribution of each parameter within the reference population. Ideally, the speech characteristics of an unseen healthy individual should fall within the RIs derived from the reference population. The concept of RIs is commonly applied in clinical laboratory science to interpret laboratory results and assess individual health.

This chapter presents our first efforts in this direction, where we explore RIs in the context of patho-

[2]Unlike a clinical disease, which has identifiable signs and symptoms, a subclinical disease lacks recognizable clinical findings. Many diseases (e.g. diabetes) often remain subclinical before manifesting clinically [Stöppler, 2021].

logical speech. First, we define a knowledge-based feature set informed by previous literature on the expected manifestations of different diseases in speech (particularly those summarized in figure 2.5, in chapter 2), including linguistic and acoustic features. A key aspect of these features is that they should be explainable and have some physical meaning, to allow meaningful discussion with the medical community. Then, we define reference intervals for these speech features using a reference population, based on CLAC corpus [Haulcy and Glass, 2021].

We then leverage the definition of reference speech for the detection of speech affecting diseases. Each disease detection task is formulated as a binary classification problem (patients versus controls). Specifically, we focus on the detection of Alzheimer's disease, using the ADReSS corpus, and Parkinson's Disease, using PC-GITA's subset of sustained vowels. We introduce various *deviation-scores* to quantify the divergence of an individual's speech from that of the reference population. These deviation-scores serve as inputs for disease classification.

The models explored to tackle the classification task are Support Vector Machines (SVM), Logistic Regression (LR), and Neural Additive Models (NAMs) [Agarwal et al., 2021]. NAMs are interpretable neural networks that provide insight into the decision process, which can be of utmost importance in the medical domain, especially to avoid the models learning spurious correlations.

The disease detection task focuses separately on Alzheimer's and Parkinson's disease due to the absence of public datasets including patients of each disease. However, this chapter aims to develop an approach extendable to multiple diseases and capable of simultaneous disease detection. Ultimately, our vision aims to propose speech as a biomarker for general health monitoring, emphasizing broader applications beyond the detection of individual diseases.

7.2 Related work

Characterizing reference speech

Some researchers have delved into characterizing certain features in the context of healthy speech, providing means and standard deviations. In particular, Teixeira and Fernandes [2014] studied jitter, shimmer, and harmonics-to-noise ratio in 34 female and 7 male speakers, focusing on sustained vowels /a/, /i/, and /u/ at various tones. No other acoustic features were included in this study. More recently, Shivkumar et al. [2020] proposed a toolkit for the extraction of clinically meaningful linguistic features, and presented statistics for these features for the healthy speakers in the AMI meeting corpus [Carletta et al., 2005]. This study, however, did not include any acoustic feature, since its purpose was to illustrate the toolkit. Hence, the aim of these two studies was not to define or comprehensively characterize reference speech.

Schwoebel et al. [2021], on the other hand, introduced the Voiceome Protocol and the corresponding Voiceome Dataset as standards to characterize reference speech. The authors reported statistics (means and standard deviations) for several acoustic and linguistic features, broken down by age range and gender, on their associated GitHub[3]. Although a valuable resource, the dataset was not made available upon request.

Reference intervals (RIs) in clinical laboratory science

Reference intervals are crucial in clinical laboratory science for interpreting quantitative pathology results, such as those from hematology tests [Jones et al., 2018]. RIs, defined by a lower and upper reference limit, represent the expected range of values in a reference population [Ozarda et al., 2018]. Laboratory results outside the RI do not necessarily imply disease but indicate the need for further medical evaluation [Ozarda et al., 2018]. There are two approaches to determine reference intervals: the direct approach and the indirect approach.

The *direct approach* refers to the traditional method that involves *selecting a reference population* of a minimum of 120 individuals for each partition (e.g. sex, age range); *collecting samples* for that population; performing statistical evaluation using non-parametric methods and outlier removal techniques; and estimating the reference interval between the two reference limits [Horowitz et al., 2010]. However, this method faces challenges such as defining health, the presence of subclinical disease, and selection bias associated with small cohorts [Jones et al., 2018]. Further guidelines can be found in [Horowitz et al., 2010]. Laboratories often report results with fewer samples than recommended, and selected individuals rarely represent overall biological variability, leading to poorly generalizable reference intervals. This limitation can be mitigated using bootstrapping [Ozarda, 2016].

An alternative method, referred to as the *indirect approach*, mines data from existing pathology databases, i.e., it is *based on laboratory results collected for other purposes*, usually for routine clinical care. These databases include results from diseased patients but also from healthy subjects, allowing for the extraction of the underlying reference distributions [Ozarda et al., 2018]. This approach is faster, cheaper, avoids patient inconvenience, and circumvents ethical issues related to informed consent from vulnerable populations [Jones et al., 2018; Ozarda, 2016], while providing extensive data for analysis. However, the presence of diseased sub-populations can influence RIs, and at least 400 subjects per partition are recommended [Jones et al., 2018]. Partitioning the population should be done according to the effects of age and gender on the results. Other co-factors, such as body-mass index, ethnicity, or collection methodology may also be considered, although this information is rarely available in routine pathology databases. Ichihara et al. [2010] describe several methods for partitioning the population.

While guidelines favor the direct approach [Horowitz et al., 2010], the indirect approach is increasingly popular, especially in pediatric and geriatric populations where data sampling is more challenging.

[3] https://github.com/jim-schwoebel/voiceome

If the underlying data distribution in the reference population is Gaussian, the reference limits that constitute the RI correspond to the $mean \pm 1.96 \times std$, in which std stands for standard deviation [Ichihara et al., 2010]. If such an assumption cannot be made, which is frequently the case for the studies of RI estimation, either data must be first transformed to a Gaussian distribution, e.g. using a Box-Cox power transformation, or a non-parametric estimation can be made. In the case of a non-parametric estimation, the limits of the RI correspond to the 2.5^{th} and 97.5^{th} percentile [Ichihara et al., 2010]. Both the non-parametric approach and the power transformation of data followed by the parametric approach provide similar results according to Ichihara et al. [2010], but the non-parametric method is recommended in some studies [Horowitz et al., 2010].

7.3 Framework overview and corpora

The analysis conducted in this chapter can be subdivided into two main tasks: Task 1 focuses on defining reference speech, and Task 2 leverages this definition to perform the detection of speech affecting diseases. Each task entailed a series of smaller steps (some of them shared across both tasks) and required specific data.

Below, we introduce the corpora required for both Task 1 and Task 2. The definition of reference speech, explored in Task 1, requires a reference population. However, some of the steps within Task 1 demand supplementary data beyond the reference population. Ideally, these additional data would originate from a distinct corpus. Nonetheless, this study utilizes control subjects from datasets employed in disease detection.

Reference Population

Aligned with the indirect approach for estimating reference intervals described in Section 7.2, the reference population used to derive reference intervals was sourced from existing publicly available databases. Since routine pathology databases typically lack speech recordings, we used CLAC Corpus [Haulcy and Glass, 2021], a corpus created to provide a collection of audio samples from healthy speakers, described in detail in chapter 3. CLAC presents several advantages. Firstly, it is substantially larger than other publicly available speech corpora in the field of speech as a biomarker for health. Secondly, it includes speech tasks frequently studied for disease detection. Thirdly, it presumably has a low incidence of unhealthy subjects. The subsets of the corpus used in this study include two picture description tasks (Cookie Theft picture and Picnic picture, typically used in the diagnosis of cognitive impairment), and a sustained vowel task (/a/). The 13 speakers who identify as “other” in terms of gender were excluded from the analysis because the sample size was too small.

Although integrating multiple datasets would be ideal, this study focuses on picture description and sustained vowel tasks to enable a comparison with publicly available datasets for AD and PD detection.

To the best of our knowledge, no other publicly available datasets include these tasks with healthy speakers, or speakers claiming to be disease-free.

Population for disease detection

Two datasets were used to explore disease detection: ADReSS [Luz et al., 2020] for the analysis of Alzheimer's disease, and PC-GITA [Orozco-Arroyave et al., 2014] for the analysis of Parkinson's disease. Both datasets have been described in chapter 3, and used in other experiments in the previous chapter. Although the audios released in the ADReSS challenge were acoustically enhanced, the experiments described in this chapter used the original version made available in the Pitt Corpus [Becker et al., 1994]. The interviewers' interventions were removed, using the annotations provided in the manual transcriptions. PC-GITA is fully spoken in Spanish, thus, to allow a fair comparison with the reference intervals determined for English, the only task explored in this work is the enunciation of a sustained vowel /a/ (three repetitions per subject).

7.4 Task 1: Reference speech characterization

Task 1 proposed an approach to characterize reference speech, by defining reference intervals for a set of knowledge-based features informed by the literature. This approach is inspired by the indirect approach for reference interval estimation, described in section 7.2.

7.4.1 Method

The task of defining reference speech started by a pre-processing step which included ASR or vowel segmentation when appropriate (step 1). This step was followed by the extraction of interpretable acoustic and linguistic features (step 2), and by the removal of outliers (step 3). Subsequently, we assessed the necessity of partitioning the reference population based on gender, age ranges, or speech tasks by employing Mann-Whitney U tests [McKnight and Najab, 2010] (step 4). The reference intervals for each feature were then established using the refined reference population (step 5). Additionally, we investigated a dimensionality reduction strategy to enhance interpretability (step 6). Figure 7.1 presents an overview of these steps.

7.4.1.A Step 1. Data pre-processing

The data pre-processing step involves the application of different sub-steps depending on the nature of the speech task. If the task involves a sustained vowel, vowel segmentation is employed. Conversely, for spontaneous speech tasks, ASR is applied.

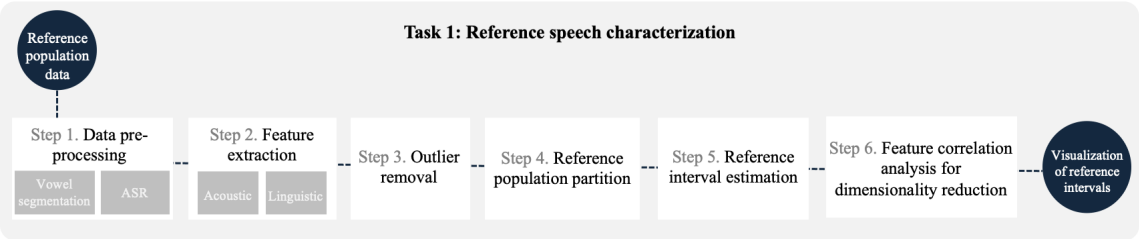


Figure 7.1: Overview of the steps entailed in task 1: Reference speech characterization.

Table 7.1: Sustained vowel pre-processing steps.

	Step	Notes
1:	Exclude any files whose maximum RMSE is below 0.005.	
2:	Search for abrupt changes in the RMSE.	Changes above 15% of the RMSE of the signal are considered abrupt changes.
3:	**IF** no abrupt changes are detected	
4:	Keep the recording.	
5:	**ELSE**	
6:	Extract the segments between each abrupt change.	
7:	Discard the segment after the last abrupt change until the end of the recording.	This segment may correspond to a gain reduction.
8:	Select the segment with the largest duration (without abrupt RMSE changes) for analysis.	
9:	**ENDIF**	
10:	Split all segments longer than 4 seconds into chunks of 3 s, with a sliding window of 2 s.	To approximate the average duration of vowel segments in PC-GITA.
11:	Before feature extraction, identify a "stable" sustained vowel segment.	A "stable" segment corresponds to at least 110 periods without voice breaks. A voice break is considered if a period is larger than the maximum phonation period, set to 0.02 s. This criteria was defined based on a speech pathologist advice.
12:	After feature extraction, exclude the segment if the standard deviation of F0 is larger than 100 Hz.	

Vowel segmentation

Due to its nature as a crowdsourced corpus, CLAC includes recordings that exhibit anomalies, particularly in the recordings of sustained vowels. Examples of these anomalies include recordings with very little energy, recordings with a decrease in gain after a couple seconds because the tool used for data collection did not recognize the sound as speech, etc. For this reason, and to improve the overall quality of the recordings that constitute the reference population, data filtering was performed to remove or segment sustained vowel recordings from CLAC, following the twelve pre-processing steps described in Table 7.1. The parameters (e.g. minimum acceptable root mean squared energy (RMSE) of 0.005, the threshold of 15% for abrupt changes, and the threshold of 100 Hz for standard deviation of F0) were empirically defined. Steps 1-10 were applied to the sustained vowel recordings in CLAC. Step 11 and 12 were applied both to the recordings in CLAC and PC-GITA, although no files exist on PC-GITA with a standard deviation of F0 larger than 100 Hz. This process resulted in the exclusion of 12 vowel recordings (out of the 300) in PC-GITA, and 355 (out of 2811) in CLAC.

Automatic speech recognition (ASR)

The extraction of linguistic features required transcriptions of the picture description task. Unlike previous

studies optimizing ASR systems for individuals with Alzheimer's and Parkinson's disease (e.g., [Hu et al., 2023]), this study adopts a zero-shot approach suitable for the general population and various speech-affecting diseases. Thus, ASR systems specifically trained on the corpora under study were avoided. Five state-of-the-art ASR systems, available at Hugging Face, were compared:

- *wavlm-libri-clean-100h-large*. This model is a fine-tuned version of Microsoft's self-supervised WavLM-Large model [Chen et al., 2022]. It has been specifically finetuned on the Librispeech ASR clean dataset, making it highly effective for recognizing clean speech data.
- *wav2vec2-large-960h* [Baevski et al., 2020]. This model developed by Facebook was pre-trained and fine-tuned on the 960 hours of Librispeech. It leverages the wav2vec 2.0 architecture, which is known for its effectiveness in self-supervised learning for speech recognition.
- *wav2vec2-large-robust-ft-swbd-300h*. This model is a version of the wav2vec2-large-robust model [Hsu et al., 2021b] fine-tuned on the Switchboard corpus [Godfrey et al., 1992], which comprises telephone speech under noisy conditions. This fine-tuning enhances its robustness to noisy recordings, making it well-suited for real-world applications where background noise is prevalent.
- *wav2vec2-large-xlsr-53-english* [Grosman, 2021]. This model is a fine-tune of Facebook's XLSR-Wav2Vec2, which learns cross-lingual speech representations by pretraining a single model from the raw waveform of speech in multiple languages. The fine-tuning was conducted in English using the train and validation splits of Common Voice 6.1 [Ardila et al., 2019].
- *whisper-large* [Radford et al., 2023]. Whisper is a general-purpose speech recognition model developed by OpenAI, based on an encoder-decoder transformer architecture. Its main advantage, and the primary reason to be used here, is that it transfers well to new domains, without any dataset-specific fine-tuning. However, it is important to note that the data used to train Whisper has not been disclosed. Therefore, there is a possibility that datasets such as ADReSS and CLAC were part of the training set.

The decision to compare multiple ASR systems, rather than using the model providing the lowest Word Error Rate (WER) on the LibriSpeech benchmark [Panayotov et al., 2015], was driven by the diverse recording conditions of CLAC and ADReSS, compared to LibriSpeech. Models trained on data with diverse recording conditions are expected to yield the best transcriptions.

Table 7.2 presents the WER of the five ASR systems on ADReSS. Performance on CLAC is not reported because manual transcriptions are not available in this corpus. Whisper achieves the best performance, likely due to its training on 680,000 hours of supervised data from the web [Radford et al., 2023]. However, Whisper often outputs transcriptions cleaner than the actual audio in terms of fluency, namely by removing fillers or repetitions, which may encode relevant information for studying cognitive

Table 7.2: Automatic speech recognition on ADReSS.

Model	WER
wavlm-libri-clean-100h-large [Chen et al., 2022]	53.7
wav2vec2-large-960h [Baevski et al., 2020]	48.0
wav2vec2-large-robust-ft-swbd-300h [Hsu et al., 2021b]	37.0
wav2vec2-large-xlsr-53-english [Grosman, 2021]	61.1
whisper-large [Radford et al., 2023]	**26.9**

Table 7.3: Examples of automatic transcriptions on ADReSS.

Example 1	
reference	(...) the mother doing **the dishes the sink** the water running (...)
whisper	(...) the mother doing **the dishes and** the water running (...)
wav2vec	(...) the mother doing **the dishes the sink** the water running (...)
Example 2	
reference	(...) and their mom and **uh well this here** (...)
whisper	(...) and their mom and **this here** (...)
wav2vec	(...) and their mom and **uhe this yeur** (...)

impairment. The second-best model, *wav2vec2-large-robust-ft-swbd-300h* (henceforth referred to as wav2vec), retains such disfluencies but sometimes produces non-existent words, potentially affecting downstream tasks. Table 7.3 provides examples of automatically generated transcriptions. Given this distinction, all subsequent analysis is made in parallel using the transcriptions generated by whisper and wav2vec.

The wav2vec ASR system failed to produce an output for 6 files in ADReSS, 3 from the train set, and 3 for the test set.

7.4.1.B Step 2. Feature Extraction

Singh [2019] distinguishes 3 processes for computational profiling of humans from their voice: knowledge-driven, data driven, or a combination of both. This work explores the latter. The mechanisms through which the different diseases impact speech, summarized in figure 2.5, motivated the definition of a knowledge-driven feature set containing 41 interpretable features. Later, feature selection was conducted using a data driven approach. This knowledge driven feature set contains both acoustic (28) and linguistic (13) features which are thoroughly described in table 7.4. The features are grouped into four categories: content-related, rhythm-related, voice quality-related and vocal tract shape-related features. Content-related features were derived from automatically generated transcriptions. While the analysis of the picture description task encompassed all features, the analysis of sustained vowels was solely based on voice-quality and vocal-tract related features.

Different methods were used to extract the features, as listed in table 7.4. Particularly the content-related features were extracted using the *BlaBla* toolkit [Shivkumar et al., 2020], dedicated scripts, or pre-trained models.

The coherence features were based on the cosine similarity between sentence embeddings of

Table 7.4: Description of the features used. Observations: In rhythm-related features, when the descriptions refer to the total time, it assumes that silences before the start and after the end of the speech signal were removed.

Category	Feature Name	Functional	Method	Description
Content	Content density	–	BlaBla	Proportion of number of open class words, i.e. nouns, verbs, adjectives and adverbs, to the number of close class words, i.e. determiners, pronouns, conjunctions and prepositions [Shivkumar et al., 2020].
	Idea density	–	BlaBla	Proportion of verbs, adjectives, adverbs, prepositions and conjunctions to all words across sentences [Shivkumar et al., 2020].
	Honoré statistic	–	BlaBla	Calculated as $(100 * log(N))/(1 - (V1)/(V))$, where V is number of unique words, $V1$ is the number of words in the vocabulary only spoken once, and N is overall text length [Shivkumar et al., 2020].
	Brunet's Index	–	BlaBla	Calculated as $N^{(V^{-0.165})}$, where V is number of unique words and N is overall text length. Measures the lexical richness. It is a version of TTR, insensitive to text-length [Shivkumar et al., 2020].
	Type-to-token ratio (TTR)	–	BlaBla	The number of word types divided by the number of word tokens [Shivkumar et al., 2020] .
	Discourse marker rate	–	BlaBla	The rate of discourse markers across all sentences [Shivkumar et al., 2020] (eg. "so, ok, anyway, right" [Cambridge University Press]).
	Polarity	–	TextBlob	Varies between $[-1, 1]$, where -1 defines a negative sentiment and 1 defines a positive sentiment.
	Repetition ratio	–	dedicated script	Number of repeated words over total number of words
	First person pronouns	–	dedicated script	Ratio of number of personal pronouns ("i", "me", "mine", "my"), to the text length.
	Coherence	mean, variability	cosine similarity	Cosine similarity between sentence embeddings of adjacent text segments (14 tokens), computed with the pretrained sentence-transformer model *all-mpnet-base-v2*. (More details in the text.)
	Coreference chain ratio	–	wl-coref	Number of coreference chains over text length.
	Ambiguous coreference chain	–	wl-coref	Number of coreference chains that start with a third-person pronoun over the number of coreference chains.
Rhythm	Speech rate	–	praat	Approximated number of syllables over total time [Feinberg, 2022].
	Articulation rate	–	praat	Approximated number of syllables over phonation time [Feinberg, 2022].
	Average syllable duration	–	praat	Average syllable duration [Feinberg, 2022].
	Mean pause duration	–	praat	Mean duration of silence segments, excluding silences before and after speech, motivated by [Weiner et al., 2016].
	Mean speech duration	–	praat	Mean duration of speech segments [Weiner et al., 2016].
	Silence rate	–	praat	Total silence time over total time, motivated by [Weiner et al., 2016].
	Silence-to-speech ratio	–	praat	Number of silent segments over the number of speech segments, motivated by [Weiner et al., 2016].
	Mean silence count	–	praat	Number of silence segments over total time, motivated by [Weiner et al., 2016]].
Voice quality	F0	mean, std	praat	Fundamental frequency of vibration of the vocal folds.
	HNR	–	praat	Compares the energy in the harmonics to the energy in the non-harmonic (noisy) components of the speech signal [Singh, 2019].
	local Jitter	–	praat	Jitter refers to cycle-to-cycle perturbations of F0 in frequency. Speech with high jitter is perceived as roughness [Singh, 2019]. Local jitter is the average absolute difference between consecutive periods, divided by the average period [Boersma and Weenink, 2024], measured in %.
	local absolute Jitter	–	praat	Average absolute difference between consecutive periods, measured in seconds [Boersma and Weenink, 2024].
	RAP Jitter	–	praat	Relative average perturbation - the average absolute difference between a period and the average of it and its two neighbours, divided by the average period [Boersma and Weenink, 2024].
	ppq5 Jitter	–	praat	Five-point Period Perturbation Quotient – same as RAP jitter but based but computed with it and its four closest neighbours [Boersma and Weenink, 2024].
	local Shimmer	–	praat	Shimmer refers to cycle-to-cycle variation of F0 in amplitude. Speech with high shimmer is perceived as buzzing [Singh, 2019]. Local shimmer is the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude [Boersma and Weenink, 2024], measured in %.
	local db Shimmer	–	praat	Average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20 [Boersma and Weenink, 2024].
	apq3 Shimmer	–	praat	Three-point Amplitude Perturbation Quotient – average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude [Boersma and Weenink, 2024].
	aqpq5 Shimmer	–	praat	Five-point Amplitude Perturbation Quotient – same as apq3, but computed with it and its four closest neighbours [Boersma and Weenink, 2024].
	apq11 Shimmer	–	praat	11-point Amplitude Perturbation Quotient – same as apq3, but computed with it and its ten closest neighbours [Boersma and Weenink, 2024].
Vocal tract	F1	mean, median	praat	Formants occur around frequencies that correspond to the resonances of the vocal tract. First formant frequency – relates to the shape of the area behind the tongue (on the throat). If the ressonator has a small area, then the formant frequency should be higher.
	F2	mean, median	praat	Second formant – relates to the shape of the area from the hump of the tongue to the tip of the lips
	F3	mean, median	praat	Third formant.
	F4	mean, median	praat	Fourth formant.

adjacent text segments, computed with the pre-trained sentence-transformer model *all-mpnet-base-v2*[4]. At the time this work was conducted, this model, trained with over 1 billion training pairs, provided the state of the art on the *Sentence Embeddings Benchmark* [Reimers and Gurevych, 2019]. The embeddings were extracted for chunks of 14 tokens. The choice of 14 tokens was rooted on two reasons: (i) in CLAC, in the task where subjects describe the cookie theft picture, the average number of words per sentence in the provided transcriptions was 15, and in the task where subjects are describing the picnic picture, the average number of words per sentence is 13 words; and (ii) according to the American Press Institute [2009], readers understand over 90% of the information when sentences have 14 words. After computing the cosine similarity of adjacent sentences, the mean and the variance are computed for the entire picture description. This measure of coherence was based on the *incoherence model*, described in [Bedi et al., 2015] and [Iter et al., 2018] for the assessment of speech of subjects suffering from psychosis and schizophrenia. The use of the variance was inspired by the concept of ongoing semantic variability, proposed by Sanz et al. [2022] as a text-level semantic marker of Alzheimer's Disease.

Ambiguous coreference chains are sequences of words or phrases in a text that refer to the same entity or concept, which start with an ambiguous pronoun. Ambiguous pronouns refer to entities not explicitly mentioned or mentioned only cataphorically, i.e., after the pronoun. The usage of ambiguous pronouns, or referential incoherence, is a common pattern in incoherent speech. The usage of ambiguous pronouns was captured following the approach of Iter et al. [2018]: (1) a pre-trained coreference resolver extracts the reference chains (i.e., the lists of terms that should refer to the same entity), and (2) if the first term in the reference chain is a third-person pronoun (he, she, they, etc.), then it is considered an ambiguous pronoun. The pre-trained coreference resolver was the wl-coref[5] [Dobrovolskii, 2021], that detained the state of the art on the CoNLL-2012 Shared Task[6] [Pradhan et al., 2012] at the time this work was conducted. The entire transcription of the picture description was used to compute this feature.

The remaining three feature categories – rhythm, voice quality, and vocal tract shape – were derived directly from the audio samples, using Praat [Boersma, 2001], through the Python package praat-parselmouth [Jadoul et al., 2018]. Praat was chosen for its frequent use in clinical practice.

For some data samples, it was not possible to extract all linguistic features. Particularly, for 9 wav2vec transcriptions in ADReSS (4 in the test set), 15 wav2vec transcriptions in CLAC, and 16 whisper transcriptions in CLAC. Feature extraction failed because the generated transcriptions were either too short (for example, did not contain more than 14 tokens to compute coherence), or no English words were recognized.

[4]available for download at https://www.sbert.net/docs/pretrained_models.html
[5]available for download at https://github.com/vdobrovolskii/wl-coref.
[6]score board available at http://nlpprogress.com/english/coreference_resolution.html.

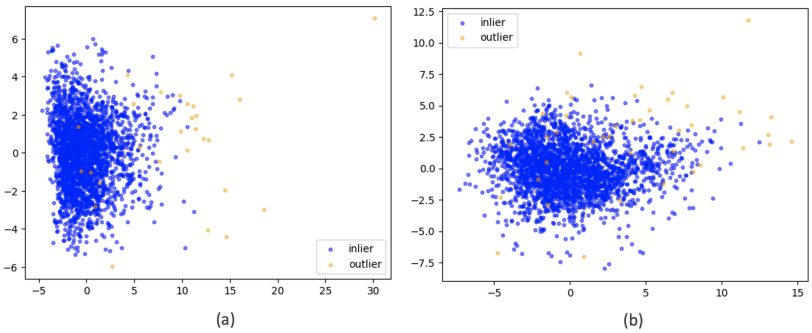


Figure 7.2: Two principal components of the reference population with identified outliers marker in yellow: (a) sustained vowel task, (b) picture description.

7.4.1.C Step 3. Outlier Removal

The identification of outliers in the reference population, i.e., samples that differ substantially from the other observations, was based on the Mahalanobis distance [Mahalanobis, 2018] to the mean of the population, an approach well-suited for multivariate data. The cutoff threshold was set to three times the standard deviation from the mean of the Mahalanobis distance. This analysis was conducted separately for the picture description and the sustained vowel tasks, using rythm, voice-quality, and vocal-tract related features. Content features were not considered to ensure that results are independent of the ASR model.

This outlier removal method excluded a total of 70 (28 vowels and 42 picture description) samples, out of the pool of 4,992 audio samples for which it was possible to extract all features (approximately 1%). By excluding these samples, we expect to exclude bad quality audios and possibly samples from speakers affected by subclinical diseases.

Figure 7.2 depicts a projection of the audio samples' features into a two-dimensional space using principal component analysis, plotting the two principal components for each (a) sustained vowel and (b) picture description sample. This visualization technique suggests that the method effectively removed the most prominent outliers. Table 7.5 presents the number of audio samples and speakers in the reference population for each speech task after outlier removal, with a detailed breakdown by gender and age range.

7.4.1.D Step 4. Reference population partition

Most features in the feature set are strongly impacted by different factors, which may include recording conditions, and speaker dependent factors, such as gender, age, body mass index, accent, education, smoking habits, etc. They may also exhibit substantial differences depending on the speech task (e.g. in spontaneous *versus* read speech). It is possible that these factors have a stronger impact on the speech signal, than that caused by speech affecting diseases, and thus could bias the results. It becomes

Table 7.5: Number of audio files and speakers, and average file duration in the reference population, per speech task, and by gender and age range, after vowel segmentation and outlier removal.

		$CLAC_{picture}$		$CLAC_{vowel}$		All	
		Files	Speakers	Files	Speakers	Files	Speakers
Count							
M	<50	1115	772	1040	598	2155	782
M	≥50	142	104	133	77	275	106
F	<50	1081	749	1044	641	2125	756
F	≥50	188	139	179	113	367	140
All		2526	1764	2396	1429	4922	1784
Average duration ± standard deviation (s)							
All		38 ± 22	–	3 ± 1	–	–	–

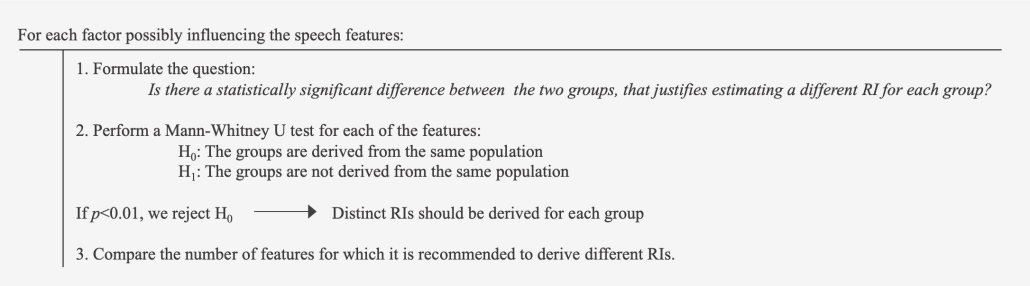


Figure 7.3: Determining whether to define different RIs for different values or ranges of each speech affecting factor.

important to determine whether different reference intervals should be estimated for different ranges of each of these factors. Due to the lack of data on all variables, this work focused on gender and age range which are the most critical factors for reference interval estimation [Ozarda, 2016]. We also explored the impact of different speech tasks and source dataset.

To determine whether the reference population should be partitioned to define different reference intervals for distinct genders, age ranges, and speech tasks, we followed the approach detailed in figure 7.3. We started by formulating the following questions:

- **Q1**: Is there a statistically significant difference between the speech features of male and female speakers, that justifies estimating different RIs for each gender group?
- **Q2**: Is there a statistically significant difference between the speech features of adult speakers at different age ranges, that justifies estimating different RIs for each age range? Here, we simplified the problem to two age ranges: subjects below 50 years old, and subjects with 50 years or more. We acknowledge that this analysis should be carried with finer age ranges, but such analysis was not possible considering the size of the reference population, particularly for the older subset.
- **Q3**: Is there a statistically significant difference between the same speech features extracted for different speech tasks?
- **Q4**: Is there a statistically significant difference between speech features of multiple datasets, that

Table 7.6: Number of features with *p-value* ≥ 0.01 in the Mann-Whitney U test applied to questions Q1-Q3. For these features, the recommendation is to derive a single reference interval valid for both subpopulations.

		Q1	Q2	Q3	Q4	Q4–dataset normalization
Total feature count		41	41	20	41	41
features with *p-value* $\geq$ 0.01	All	11	31	–		
	Female	–	31	4	23	39
	Male	–	31	4	18	40
	F & M	–	26	2	14	39

would impede combining datasets under the same RI?

To answer these questions, we performed a Mann-Whitney U test [McKnight and Najab, 2010] that evaluates whether two subgroups are likely to be derived from the same population (null hypothesis). A *p-value* $>= 0.01$ indicates that there is not sufficient evidence to reject the null hypothesis, i.e., there is not sufficient evidence supporting that the two subgroups belong to different distributions, allowing us to assume they can be represented under the same reference interval. Table 7.6 summarizes the number of features with a *p-value* $>= 0.01$, for each question. The findings presented refer to an analysis conducted using the 28 acoustic features, and 13 linguistic features extracted from whisper transcripts. A similar analysis could be reported for wav2vec transcriptions. A detailed description follows below.

To answer Q1, we analysed data for a single speech task from a single dataset, to minimize other potential sources of variability: *CLAC–picture_description*. We computed the Mann-Whitney U test between male and female individuals, and concluded that for 11 out of the 41 features there is no evidence supporting partitioning the reference population and deriving distinct RI. These 11 features are mostly linguistic (7), but also rhythm-related (4). Given that for the majority of the features, the recommendation is to partition, henceforth we developed two separate models: one for each gender.

Regarding Q2, we conduct the analysis both before and after partitioning the reference population by gender. Our analysis prior to gender-based partition indicates that there is no statistically significant difference between the features derived for adults younger and older than 50 years old for 31 out of the 41 features. Similarly, upon analysing the results posterior to gender-based partitioning, we identified 31 features (out of the 41) for which no partitioning by age is required, for both female and male subpopulations. However, these 31 features are not identical for both genders. Specifically, there is an overlap of 26 features where it is not advised to partition by age. The features for which it is recommended to derive distinct RIs per age range are mostly voice quality related features (9), but also linguistic (3), vocal tract (2) and rhythm (1). Although we have reached a recommendation to partition the population by age range for 10 out of the 41 features, both when analysing the entire reference population, and in each gender subgroup, in this initial work, we opted to not partition the data according to age. This choice was motivated by the fact that partitioning by age after partitioning by gender would result in small sample

sizes for each group (often much smaller than the 400 subjects suggested in the guidelines).

To investigate Q3, we compared *CLAC–picture_description* and *CLAC–vowel_a*. This investigation focused exclusively on voice quality and vocal tract shape features, as linguistic and rhythmic features are not well-suited for the analysis of sustained vowels. Our findings revealed that, within the male and female subgroups, only four features could be consistently analyzed together for the two speech tasks. When combining results across both genders, only two features remained viable. Consequently, the analyses of different speech tasks should be conducted separately. In other experiments [Botelho et al., 2023], we compared two speech tasks of more similar nature – picture description task and read speech task – and reached the same conclusions.

Regarding Q4, we compared *CLAC–picture_description* and the controls in ADReSS. In this scenario, we could compute the RIs using all data together for only 14 out of the 41 features. This limitation implies that the RIs estimated with one dataset may not generalize to unseen datasets, even when considering the same gender and speech task. To tackle this issue, similarly to the approach in section 6.4, we performed dataset-dependent normalization, specifically zero-mean and unit variance normalization. With this approach, we made the distributions of the two subgroups more similar, and it can be assumed that the samples come from the same distribution for 39 out of the 41 features.

The two features for which we could not assume the same distribution between datasets were the *ratio of ambiguous coreference chains* and *discourse marker rate*. However, it is noteworthy that for a similar analysis using wav2vec transcriptions instead of whisper transcriptions (omitted here to maintain the focus of the chapter), the feature *discourse marker rate* was not included in this list. This difference underscores how whisper transcriptions interfere with extraction of filler related information.

In summary, we conclude that we need to estimate different RIs for each gender, and each speech task. To combine different datasets under the same RI, we should also perform dataset-dependent normalization. This is consistent with results described in [Botelho et al., 2020b].

These are simplifying assumptions, that we believe to be reasonable in this proof-of-concept exploring the feasibility of defining RIs for speech. Future work should not only study a larger reference population, but also consider other methods for partitioning the RIs, such as the Lahti criteria [Lahti et al., 2002], or the Ichiahara method [Ichihara et al., 2008, 2010].

7.4.1.E Step 5. Reference intervals estimation

In this work, an RI, i.e., the interval between the 2.5^{th} and 97.5^{th} percentiles [Ichihara et al., 2010], is derived for each feature, using the non-parametric approach. Following the guidelines in [Ozarda, 2016], 90% confidence intervals (CIs) were derived for both the lower and upper limits of the RI via boostrapping [Ferrer and Riera], to provide a confidence measure on the estimated RI. Data was resampled 1000 times to estimate the confidence interval.

We acknowledge that for certain features, it is more appropriate to provide a single boundary, either an upper or lower limit. For instance, elevated values of jitter and shimmer are considered pathological, possibly indicating affected laryngeal control, whereas there is no lower limit below which these values are deemed unhealthy. We believe that determining whether each feature should have a reference interval or a single limit should be guided by domain-specific knowledge, and we encourage further research on this topic. Our data-driven approach does not allow us to draw conclusions on this matter; therefore, we will derive reference intervals with two limits for all features.

7.4.1.F Step 6. Feature correlation analysis

Some of the features in the proposed feature set are very correlated with each other. It is expected, for instance, that the mean and median of the formants are very correlated with each other, and also measures of shimmer, and measures of jitter should be correlated amongst each other. Additionally, a high dimensionality feature space hinders the interpretability of the results by the medical community.

Therefore, we carried out a feature correlation analysis to exclude redundant features and reduce the dimensionality. Features were grouped into clusters of similar information using hierarchical clustering based on the Pearson correlation between all feature pairs. Different correlation thresholds, $CT \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, were explored to fix the final clusters. For $CT = 1$, each cluster corresponded to a single feature. For other correlation thresholds, a *"prototype feature"* was selected to represent each cluster, resulting in the final reduced-dimensionality feature set.

To promote robustness to dataset shifts, prototype features were chosen based on the similarity of the standard deviation of their distributions across different datasets. Ideally, other corpora of control individuals would be available for this analysis. Since that is not the case, the standard deviation of each feature in the reference population was compared to that of control subjects in the disease detection population. The feature with the most similar standard deviation in both groups within each cluster was designated the prototype feature. Means were not considered, as they can be adjusted by adding a bias term. Future work should explore more sophisticated methods for selecting prototype features.

This dimensionality reduction approach, which involves Pearson correlation analysis followed by hierarchical clustering on the reference population, was motivated by two primary objectives: first, to establish a feature set suitable for characterizing reference speech independently of disease-specific deviations; second, to mitigate the risks of unstable results, overfitting, and poor generalization associated with supervised feature selection on small datasets [Dernoncourt et al., 2014; Soares et al., 2016; Vabalas et al., 2019].

7.4.2 Results

Reference intervals were derived for each feature in the full-dimensionality feature set[7], with the corresponding confidence intervals on the upper and lower bound. The analysis was conducted separately for the sustained vowel /a/ and the picture description task. For the content related features, the intervals were derived separately for each ASR method – whisper and wav2vec.

7.4.2.A Feature correlation analysis

As described in section 7.4.1.F, a reduced-dimensionality feature set was derived to exclude features highly correlated with each other. Figure 7.4 (a) shows a heatmap with the Pearson correlation values between all feature pairs. Strong colored cells correspond to pairs of strongly correlated features – red for positively correlated features, and blue for inversely correlated features. The values above the diagonal correspond to features derived for female subjects, while the values below the diagonal correspond to values derived for male subjects. As expected, one can observe that mean and median values of formant frequencies are strongly correlated, as well as the different measures of jitter and different measures of shimmer. HNR also appears inversely correlated with shimmer measures. Other patterns appear, for example speech rate is inversely correlated with silence rate, average syllable duration is inversely correlated with the articulation rate, and the repetition ratio is positively correlated with the Brunet's Index and inversely correlated with the type-to-token ratio. Correlations between features are similar for both female and male subjects. The clusters of highly correlated features, achieved via hierarchical clustering, are represented in Figure 7.4 (b). The *prototype-features* of each cluster, i.e., the features that were selected to represent the information encoded in each cluster, as described in section 7.4.1.F, are listed in table 7.7. Naturally, the higher the correlation threshold, the larger the number of clusters, and thus the higher the dimensionality of the final set.

7.4.2.B Qualitative comparison of reference speech and patient speech

Figures 7.5 and 7.6 show reference intervals represented in a radar chart, normalized to zero mean and unit variance, for the sustained vowel task and picture description task, respectively.

In the top plots, light green lines indicate the lower and upper bound of the reference interval, with the shaded region representing the confidence interval on the reference interval limits. The dark green line represents the mean values, which are always zero due to normalization. The first and third columns of each figure display the reference intervals for the entire feature set, while the second and fourth columns

[7] Although the feature *ratio of ambiguous coreference chains* may be interesting for the detection of several diseases, including Alzheimer's disease and schizophrenia, it was excluded from further analysis for the following reasons: (1) the findings in section 7.4.1.D suggest it is not robust to dataset shifts, not even with stratified normalization, and (2) the confidence intervals on the upper bound for both genders and both ASR systems were larger than the RI itself, which indicates a poor confidence on the derived RI. Further discussion on this feature is provided in Appendix E.

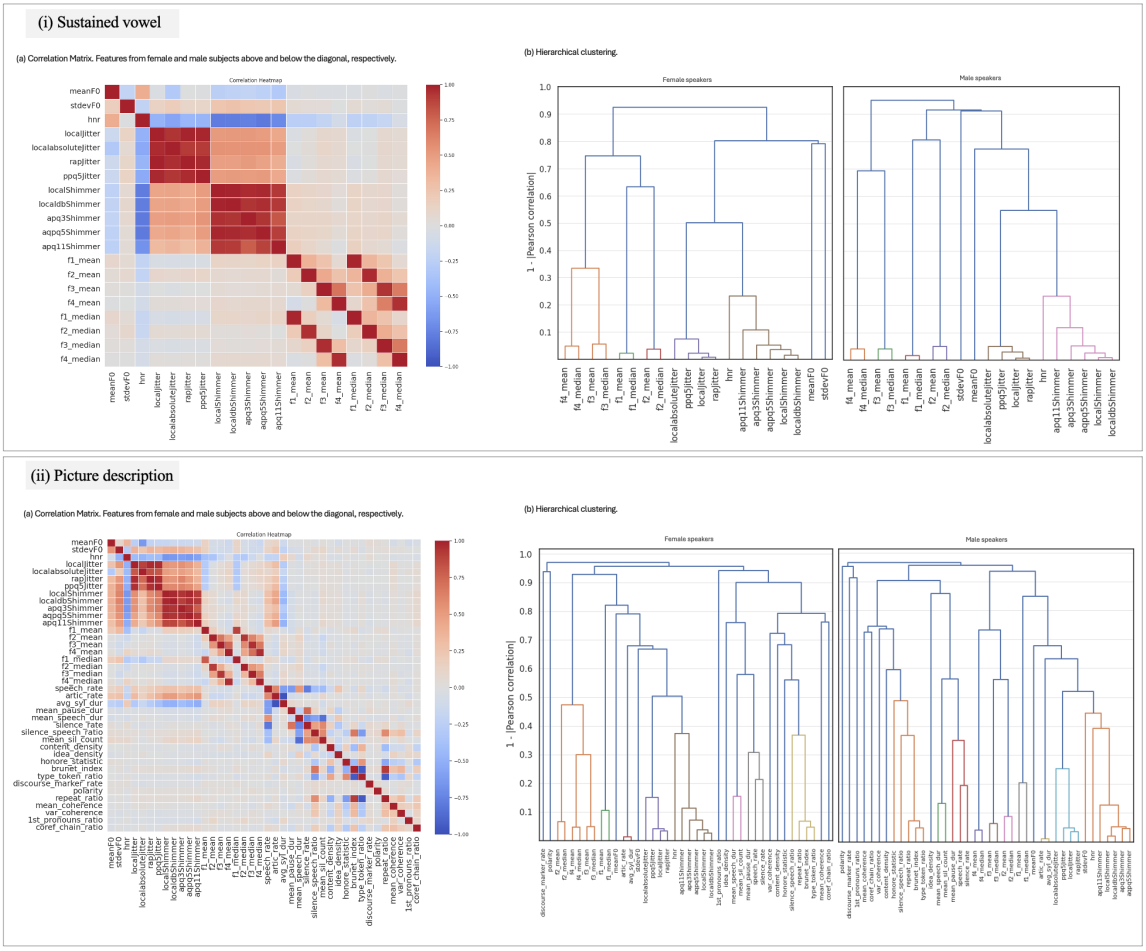


Figure 7.4: Correlation analysis of the features extracted from vowel recordings (top) and picture description using whisper transcriptions (bottom). (a) shows the Pearson correlation between the features. The values above the diagonal refer to features extracted for female subjects, while the values below the diagonal refer to male subjects. (b) shows the dendograms results from the hierarchical clustering of features, based on their Pearson correlation correlation. The y axis corresponds $1 - CT$, to capture the distance between features of the same cluster.

refer to the feature subset obtained after the Pearson correlation-based feature selection, with $CT = 0.5$. This reduced, less correlated, feature set aims to highlight which groups of features are more impacted by each disease. The speech of any subject, while performing one of the speech tasks analysed, can be projected into the radar plot, and compared to the reference population. Ideally, if a subject is healthy, their speech should be represented within the area delimited by the reference intervals.

On the second and third-row plots, we overlay individual data from the disease detection population onto the reference interval radar charts. Following the discussion on Q4 in section 7.4.1.D, the population for disease detection was also normalized to zero-mean and unit variance, using only the control subjects to compute the statistics for normalization. Each subject is represented by a different line, control subjects in blue and patients in magenta, shown in separate plots. The entire area within the reference interval is shaded to enhance visibility.

Table 7.7: *Prototype-features*, per correlation threshold, CT.

CT	Champion-features
Sustained vowel	
0.5 / **0.6** / **0.7**	F1_median, F2_mean, F3_mean, F4_mean, HNR, localabsoluteJitter, meanF0, stdevF0
0.8	apq11Shimmer, aqpq5Shimmer, F1_median, F2_mean, F3_mean, F4_mean, HNR, localabsoluteJitter, meanF0, stdevF0
0.9	apq11Shimmer, aqpq5Shimmer, F1_median, F2_mean, F3_mean, F4_mean, HNR, localabsoluteJitter, localdbShimmer, meanF0, stdevF0
Picture description (whisper transcriptions)	
0.5 / **0.6**	First person pronouns, articulation rate, content density, discourse marker rate, F1_mean, F2_median, F3_mean, F4_median, HNR, Honoré statistic, idea density, localabsoluteJitter, meanF0, mean coherence, mean silence count, polarity, ppq5Jitter, coreference chain ratio, repetition ratio, speech rate, stdevF0, coherence variability
0.7	First person pronouns, articulation rate, content density, discourse marker rate, F1_mean, F2_median, F3_mean, F4_mean, F4_median, HNR, Honoré statistic, idea density, localabsoluteJitter, localdbShimmer, meanF0, mean coherence, mean pause duration, mean silence count, polarity, ppq5Jitter, coreference chain ratio, repetition ratio, silence-to-speech ratio, speech rate, stdevF0, TTR, coherence variability
0.8	First person pronouns, articulation rate, content density, discourse marker rate, F1_mean, F1_median, F2_median, F3_mean, F4_mean, F4_median, HNR, Honoré statistic, idea density, localabsoluteJitter, localdbShimmer, meanF0, mean coherence, mean pause duration, mean silence count, polarity, ppq5Jitter, coreference chain ratio, repetition ratio, silence rate, silence-to-speech ratio, speech rate, stdevF0, TTR, coherence variability
0.9	First person pronouns, apq11Shimmer, articulation rate, content density, discourse marker rate, F1_mean, F1_median, F2_median, F3_mean, F4_mean, F4_median, HNR, Honoré statistic, idea density, localabsoluteJitter, localdbShimmer, meanF0, mean coherence, mean pause duration, mean silence count, mean speech duration, polarity, ppq5Jitter, coreference chain ratio, repetition ratio, silence rate, silence-to-speech ratio, speech rate, stdevF0, TTR, coherence variability

For the picture description task, the figure illustrates only the content features generated using Whisper transcriptions as an example. A similar plot could be derived using wav2vec transcriptions.

By visual inspection, the confidence intervals on the limits of the RIs derived for the sustained vowel task (Figure 7.5) appear relatively narrow, with the exception of the *standard deviation of F0* and the *mean of the second formant* for male subjects. For the picture description task (Figure 7.6), the feature with a wider confidence interval is *silence-to-speech ratio*. One can interpret these wide confidence intervals as an indication of lack of confidence on the exact margins of the RI. Future research should aim at improving the confidence of these intervals with a larger reference population, collected under controlled recording conditions.

When visually inspecting the plots representing the PD patients enunciating a sustained vowel /a/ (Figure 7.5), it is clear that they deviate from the reference values more frequently than controls in the axis that correspond to HNR, jitter, shimmer and F0 related features. This is notorious for both genders, however it appears that HNR, jitter and shimmer features are more relevant for female subjects, while F0 features are more relevant for male subjects.

In analyzing the Cookie Theft image description, differences between AD patients and controls also exist (Figure 7.6). For instance, several female AD patients exhibit a discourse marker rate (feature 33) substantially above the reference interval. Specifically, 26% of female AD patients surpass the RI for discourse marker rate, compared to only 7% of female controls. Additionally, the speech rate (feature 20) for both male and female AD patients is more frequently below the RI than that of control subjects.

However, these differences appear more notorious in the sustained vowel task than in the picture description. This may indicate that the task of enunciating a sustained vowel may be more suitable for this RI analysis, as it entails less sources of variability. It is also possible that the noisy recording

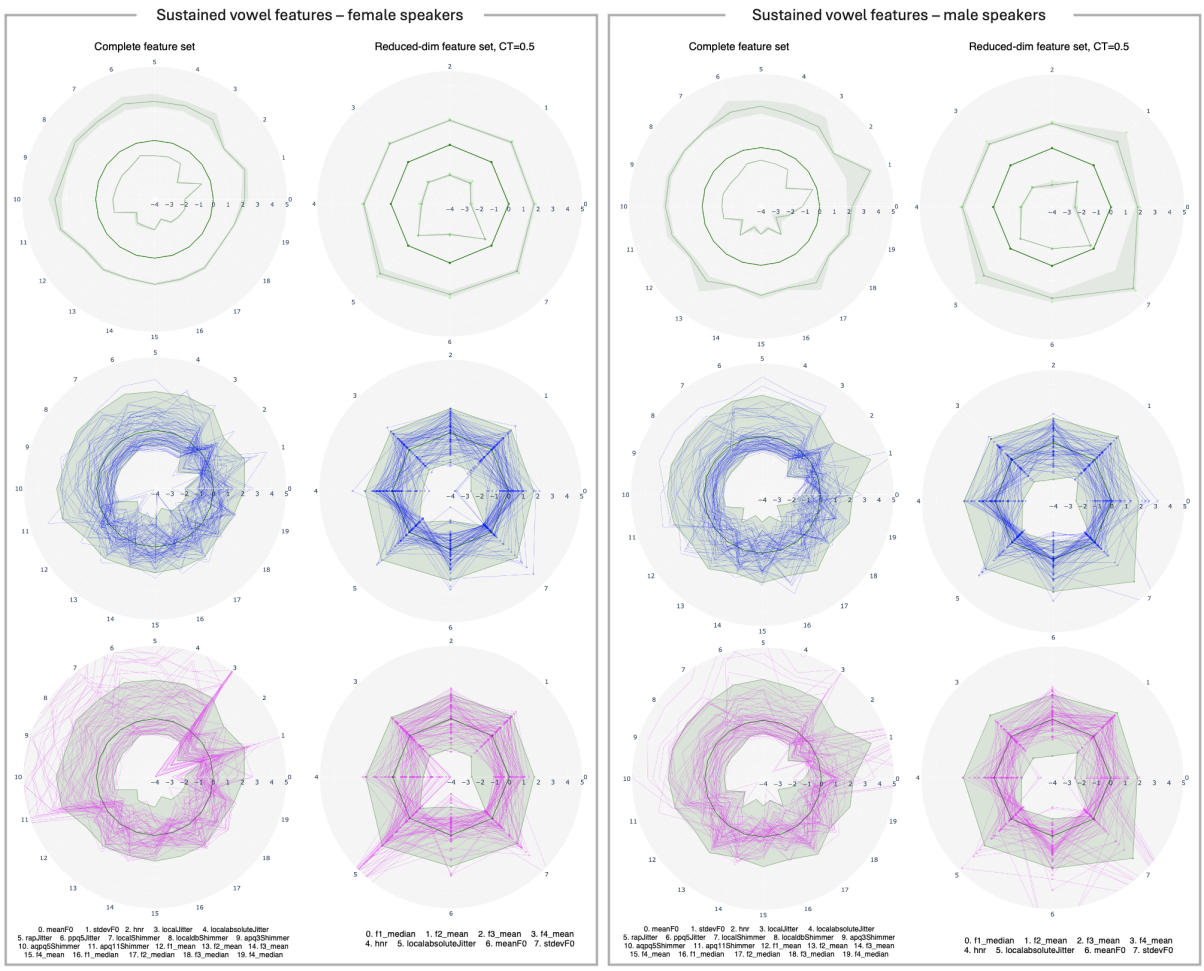


Figure 7.5: Radar plots to characterize reference speech, using the task *sustained vowel /a/*. The dark green line corresponds to the mean value of each feature, while the light green lines correspond to the reference interval, computed using the reference population. Blue lines correspond to control speakers, whereas pink lines correspond to patients (PD).

conditions in DementiaBank play a strong role.

This radar chart visualization is particularly well-suited for the analysis of speech as a biomarker for health in two scenarios. When studying a disease population, the radar chart visualization enables the identification of features that appear to be strong markers of a disease, and simultaneously still robust to dataset shifts. Taking the example of PD female patients *vs* Control females (Figure 7.5 – left): there are 6 features (all jitter-, shimmer- and HNR-related) for which more than 95% of the controls stay inside the RI and over 20% of the patients fall outside the RI. Alternatively, this visualization provides a simple way to compare the speech features of one individual to the speech features of a reference population, and quickly identify if there are any deviations on groups of features that are expected to be affected by a certain disease.

Radar plots have been previously employed to visualize speech features in the context of speech

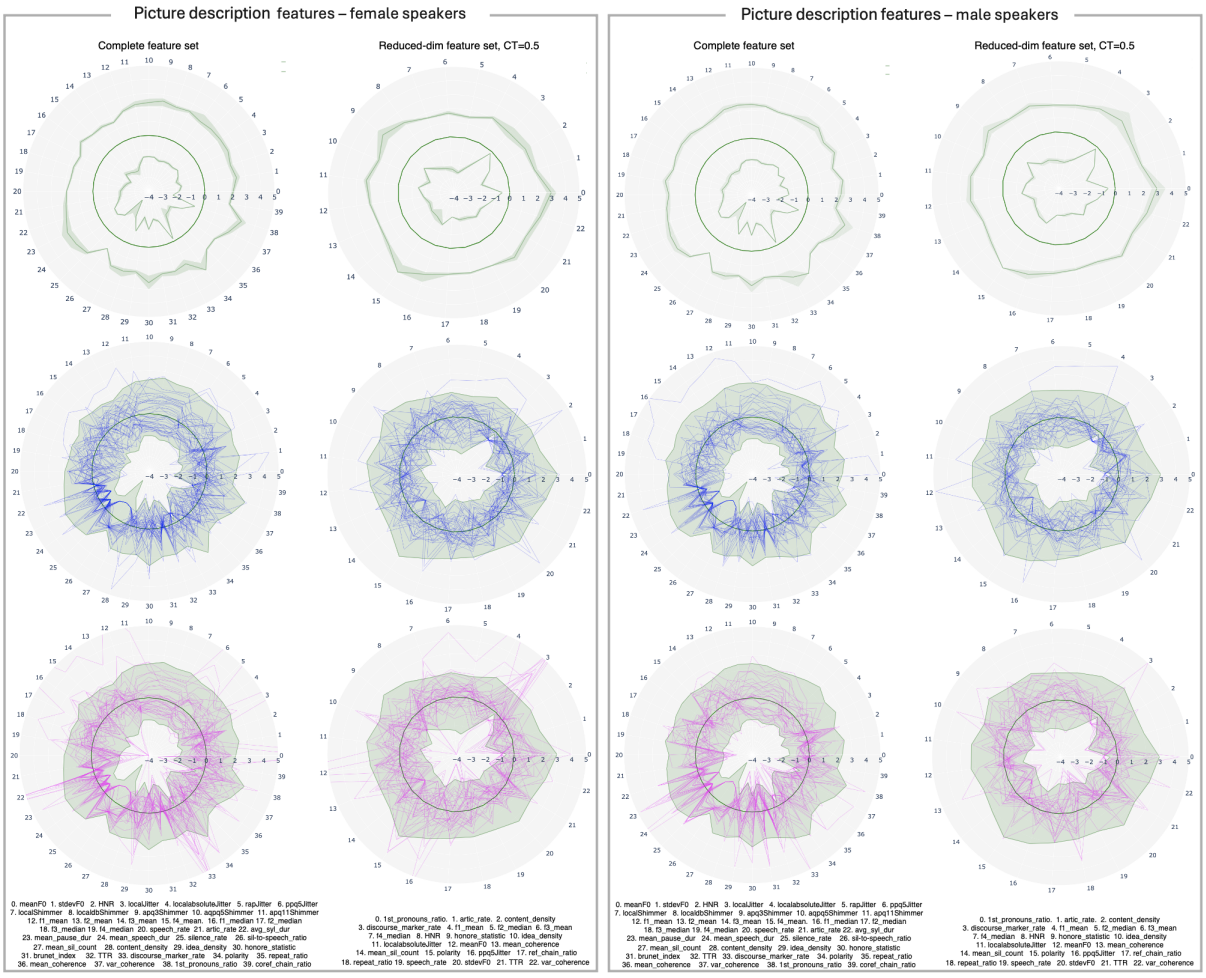


Figure 7.6: Radar plots to characterize reference speech, using the task *picture description*, and whisper transcriptions. The dark green line corresponds to the mean value of each feature, while the light green lines correspond to the reference interval, computed using the reference population. Blue lines correspond to control speakers, whereas pink lines correspond to patients (AD).

as a biomarker. Jiao et al. [2017] used radar plots to illustrate phonological disturbances in dysarthric speakers, while Behrendt [2023] introduced DemVis, a prototype system for extracting and visualizing speech features, which also performs AD detection.

7.4.2.C Quantitative comparison of reference speech and disease-affected speech

Table 7.8 reports (i) the number of samples under analysis, (ii) the average number of features per audio sample that fall outside the RI, and (iii) the average distance of features from the RI limit, calculated per audio sample. The distance was computed as the difference between the feature value and the RI margin, divided by the length of the interval. If the feature value lies within the RI, the distance is considered 0.

The Table indicates that patients have a higher average number of features outside the RI per sample

Table 7.8: Results of the quantitative comparison between reference speech and disease-affected speech. (i) Number of audio samples. (ii) Average number of features outside of the RIs, per audio sample. (iii) Average feature distance to the limit of the RI, per audio sample. Because we could not assume the values follow a Normal distribution, *p-values* were computed with a Mann-Whitney U test. (*) indicates statistically significant differences between controls (C) and patients (P).

	(i) # Samples				(ii) # Features outside of RI						(iii) Feature distance to the RI					
	Female		Male		Female			Male			Female			Male		
	C	P	C	P	C	P	*p-value*	C	P	*p-value*	C	P	*p-value*	C	P	*p-value*
PC-GITA	75	74	70	69	1.5	3.5	0.0092*	1.4	2.0	0.327	0.006	0.131	0.0003*	0.008	0.038	0.0392*
ADReSS	43	43	35	35	3.0	4.8	0.0034*	3.3	4.6	0.004*	0.014	0.050	0.0001*	0.013	0.037	0.0005*

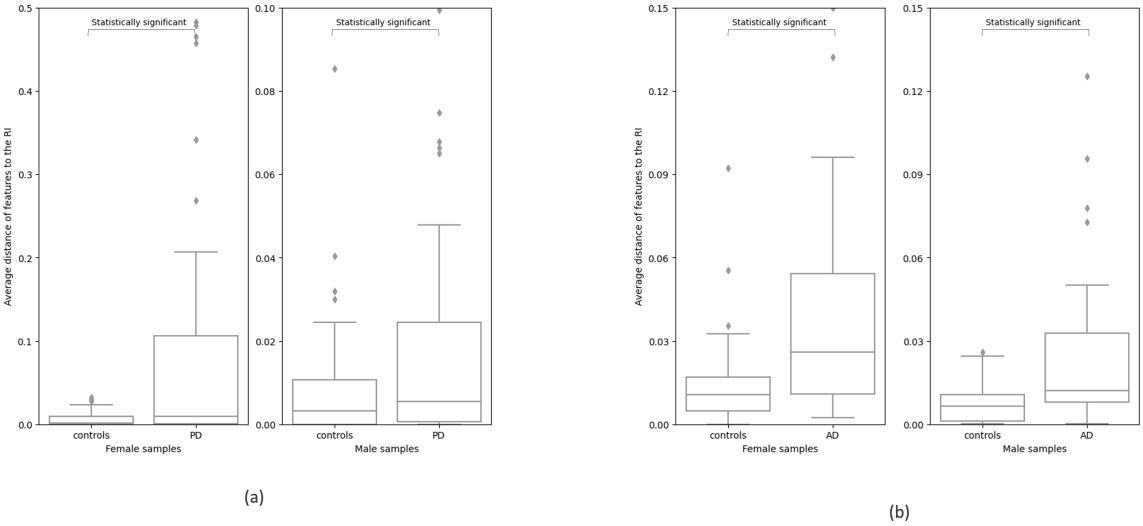


Figure 7.7: Distribution of the average distance of the features to the RI limits, per audio sample. (a) represents PC-GITA, and (b) represents ADReSS.

compared to control subjects. This difference is statistically significant in all cases, except for male speakers in PC-GITA.

More relevant than the number of features outside the reference interval, is the distance of these features from the interval. Table 7.8 (iii) reports the average distance per group, and Figure 7.7 illustrates the distribution of these values. This distance is, on average, higher for patients, and the difference between the two groups is considered statistically significant for all cases.

Future work could analyse this average distance per group of features considered relevant for each disease.

7.5 Task 2: Classification of multiple speech affecting diseases

Task 2 leverages the definition of reference speech for the automatic detection of speech-affecting diseases. Specifically, we introduced *deviation-scores* to quantify the extent to which the speech features of each new individual deviate from the reference values of the reference population. These deviation

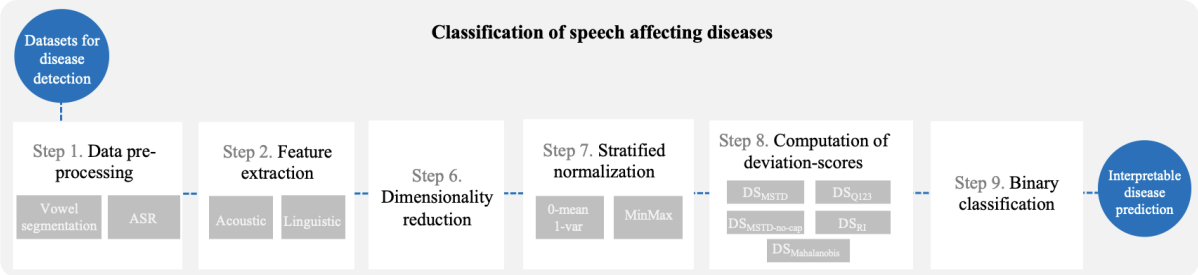


Figure 7.8: Overview of the steps entailed in task 2: Detection of diseases.

scores served as inputs for binary classifiers, including Support Vector Machines, Logistic Regression, and Neural Additive Models (NAMs). NAMs offer the advantage of full transparency, enabling explanations that are compatible with clinical reasoning.

This analysis focuses on the detection of Alzheimer's disease and Parkinson's disease separately, but ultimately it is aimed at defining an approach suitable for multiple speech-affecting diseases. In fact, we propose a single framework that could be used for the detection of several diseases, differing slightly on the subsets of features to be used, depending on the speech task at hand, and on the model used for classification. As described earlier in section 7.4.1.B, for the picture description task, all feature groups should be used, while for sustained vowels, only *voice quality* and *vocal tract* related features are suitable. We expect that future work analysing reading tasks should include all feature groups, except content-related features, and spontaneous speech tasks should leverage all feature groups.

7.5.1 Method

Task 2 entails several steps, some of them – pre-processing, feature extraction, and dimensionality reduction – shared with Task, as depicted in Figures 7.1 and 7.8. As described in sections 7.4.1.A, and 7.4.1.B, the data pre-processing and feature extraction steps resulted in the exclusion of some samples from ADReSS and PC-GITA. These samples were not used for classifier training or hyper-parameter optimization. However, to ensure a fair comparison with previous literature, these excluded samples are arbitrarily assigned the prediction "control" when reporting performance on the test set. This reflects a 50% a priori probability of a correct prediction due to the balanced nature of the datasets.

After data pre-processing, feature extraction, and dimensionality reduction, data underwent normalization (step 7), and deviation scores were computed (step 8). This process culminated in binary classification (Step 9). Detailed descriptions of these steps are provided below. While steps 3 to 5 pertain exclusively to the speech characterization pipeline (Task 1) and are not part of the disease detection pipeline (Task 2), the numbering is retained across both pipelines for consistency, reflecting the shared nature of some steps.

7.5.1.A Step 7. Data normalization

Prior to comparing the datasets for disease detection with the reference intervals and calculating deviation scores, stratified normalization by gender and source dataset was performed. This approach preserves the intrinsic characteristics of each group, acknowledging that the relationship between controls and patients may differ between male and female subjects, and that the distribution of each feature may differ in each dataset. Only the control subjects in the training set within each stratification group were used to compute the statistics for normalization. We compared two normalization strategies, zero mean and unit variance scaling, and *MinMax* scaling between zero and one. Both strategies were implemented using the scikit-learn toolkit [Pedregosa et al., 2011]. This approach assumes the gender of speakers in the test set is known, a reasonable assumption given the high performance of automatic speech-based gender detection methods [Kwasny and Hemmerling, 2021].

7.5.1.B Step 8. Deviation-scores computation

In the previous task, we characterized reference speech using the distribution of acoustic and linguistic features within a reference population. The hypothesis explored here is that deviations of a new audio sample relative to the reference population can indicate the presence of a specific disease in the speaker. Five deviation scores (DS) were compared to assess the extent to which each feature value x_i in a new individual diverges from the corresponding feature distribution in the reference population.

1. DS_{MSTD}: This deviation-score, inspired by [Zusag et al., 2023], is based on the mean, μ, and standard deviation, σ, of the feature distributions within the reference population, and it is computed as $DS_{MSTD_i} = 1 - \frac{\sigma_i}{|\mu_i - x_i|}$ if $|\mu_i - x_i| > \sigma_i$ else it is set to 0.

2. $DS_{MSTD-no-cap}$: This deviation-score is similar to the previous one, except it is not capped at 0 when the feature values are inside the interval $[\mu_i - \sigma_i, \mu_i + \sigma_i]$. The idea with this deviation-score was to approximate it to the log-likelihood ratio. It is computed as: $DS_{MSTD-no-cap_i} = \frac{|\mu_i - x_i|}{\sigma_i}$.

3. DS_{Q123}: This deviation-score is proposed as an alternative to DS_{MSTD}, that is based on the median (Q_2) and the first and third quartiles (Q_1 and Q_3), instead of the mean and standard deviation. A second modification introduced in this deviation-score is that it yields negative or positive scores depending on whether the feature values fall below or above the interval $[Q_1, Q_3]$. This approach reflects the intuition that for certain features, deviating below or above the normal range does not carry the same implications, as discussed earlier in section 7.4.1.E. DS_{Q123} is

computed as:

$$DS_{Q123_i} = \begin{cases} \frac{2|Q_{3i}-x_i|}{|Q_{3i}-Q_{1i}|} & \text{if } x_i > Q_{3i}, \\ -\frac{2|Q_{1i}-x_i|}{|Q_{3i}-Q_{1i}|} & \text{if } x_i < Q_{1i}, \\ 0 & \text{elsewhere.} \end{cases} \tag{7.1}$$

4. DS_{RI}: This deviation-score yields the same score for all feature values inside the reference interval, i.e., between the lower and upper bound of the reference interval, $[RI_{LB}, RI_{UB}]$. It also yields negative values for feature values below the reference interval. It is computed as:

$$DS_{RI_i} = \begin{cases} \frac{2|RI_{UBi}-x_i|}{|RI_{UBi}-RI_{LBi}|} & \text{if } x_i > RI_{UBi}, \\ -\frac{2|RI_{LBi}-x_i|}{|RI_{UBi}-RI_{LBi}|} & \text{if } x_i < RI_{LBi}, \\ 0 & \text{elsewhere.} \end{cases} \tag{7.2}$$

5. $DS_{Mahalanobis}$: This deviation-score consists of computing the Mahalanobis distance of each new audio sample to the median (Q_2) of the reference population. Unlike the other deviation scores that are computed at the feature level, this deviation score is multivariate and considers the deviation of the vector of all features, x to the reference population. Thus, it provides a single value for each audio sample. It is computed as follows: $DS_{Mahalanobis} = \sqrt{(x-Q_2)V^{-1}(x-Q_2)^T}$, where V is the covariance computed on the reference population. The Mahalanobis distance has been employed in [Jiao et al., 2017] to compute the distance of a dysarthric speaker from the healthy distribution, based on phonological features.

7.5.1.C Step 9. Binary Classification

The classification task was performed using three classifiers: Support Vector Machines (SVM), Logistic Regression (LR), and Neural Additive Model (NAM). The binary classification experiments were based on ADReSS for Alzheimer's disease detection, and PC-GITA for Parkinson's disease detection. Given the limited size of these corpora, the experiments were conducted in a 10-fold cross validation (CV) setting. For ADReSS, because a held-out test set was defined for the challenge in which the corpus was introduced, the 10-fold cross validation was applied on the training set, for hyperparameter tuning. Afterwards, the 10 models trained during CV each make predictions for the test set, and these predictions are aggregated via majority voting. For PC-GITA, there was no held out test set, hence the 10-fold CV was conducted on the entire dataset. In each run, one of the 9 training folds was assigned as development fold, to perform hyperparameter tuning. Folds were defined to ensure that all data from the same speaker is assigned to the same fold, to avoid leakage of speaker information across training, development and test folds. Folds ensure a balance between healthy controls and patients, in terms of number of speakers, gender and age.

Data was normalized separately for both genders, as described in section 7.5.1.A, with the normalization statistics being computed based on the controls in the training folds. Distance scores were also computed separately for both genders, given that the reference intervals were derived separately for both genders. The classifiers, however, are gender-independent.

Classification experiments with Support Vector Machines and Logistic Regression

For the SVM and LR classification experiments, two transcription types (whisper and wav2vec), three normalization strategies (zero-mean and unit variance, *MinMax*, and no normalization), six correlation threshold values, and five deviation-scores were compared. To evaluate whether the deviation scores provide an advantage over using directly the features as input to the classifiers, two extra scenarios were also considered: one where the features are directly fed to the classifiers after the stratified normalization strategy described in section 7.5.1.A, and another where the features are fed to the classifier after the regular approach of normalizing all samples in the training folds together.

The SVM hyperparameters were chosen based on a grid search on the development folds, although on a relatively small parameter space, to avoid getting high complexity models, which are more prone to overfitting. The hyperparameters compared were: $kernel \in \{linear, rbf, poly\}$, $C \in \{0.01, 0.1, 1\}$, and $degree \in \{2, 3\}$.

Classification experiments with Neural Additive Models

Neural Additive Models (NAMs) [Agarwal et al., 2021] are a type of glass-box models inherently interpretable. NAMs are part of the model family called Generalized Additive Models (GAMs), which are described by

$$g(\mathbb{E}[y]) = \beta + f_1(x_1) + f_2(x_2) + \cdots + f_K(x_K), \tag{7.3}$$

where x$= (x_1, x_2, ..., x_K)$ is the input with K features, y is the target variable, g is the link function, and f_i is a univariate shape function with $\mathbb{E}[f_i] = 0$.

The idea of NAMs is to parameterize each f_i in Equation 7.3 by a neural network (subnet). In short, NAMs are a linear combination of neural networks, each attending to a single feature, that are trained jointly using backpropagation. It is this modularity that makes NAMs' predictions very easily interpretable. NAMs' predictions can be interpreted by visualizing each of the learned shape functions, e.g. plotting $f_i(x_i)$ *vs* x_i. The graphs learnt by NAMs are not *a posteriori* explanations, but rather an exact description of how the model comes to a prediction.

The NAM architecture is also compatible with a multitask-scenario, particularly suitable for the simultaneous detection of multiple speech affecting diseases, when adequate data is available. The multitask architecture is identical to that of single task NAM except that each feature is associated with multiple subnets and the model jointly learns a task-specific weighted sum over their outputs that determines the

Table 7.9: Best disease classification results, using SVM and logistic regression, in terms of accuracy (Acc), macro precision (P), macro recall (R), and macro F1, in [%].

					Dev Folds				Test				Test – Speaker MV			
CT	**DS**	**norm**	**classifier**	**ASR**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**
Parkinson's Disease																
1	DS_{RI}	MinMax	LR	–	72.2	72.7	72.2	72.1	71.7	72.2	71.7	71.5	75.0	75.5	75.0	74.9
0.9	DS_{Q123}	MinMax	SVM (linear, $C = 0.01$)	–	69.4	69.9	69.4	69.2	69.7	70.2	69.7	69.5	77.0	77.9	77.0	76.8
Alzheimer's Disease																
1.0	Features	MinMax	SVM (poly, $C = 1$, $d = 2$)	Whisper	75.9	76.8	75.9	75.7	68.8	68.8	68.8	68.7	–	–	–	–
0.7	DS_{RI}	None	LR	Whisper	70.4	71.8	70.4	69.9	77.1	77.5	77.1	77.0	–	–	–	–

shape function for each feature and task [Agarwal et al., 2021].

The combinations of ASR, normalization, deviation-scores and correlation threshold that yielded the best results using SVM and LR were used to conduct the NAM experiments. Hyperparameter tuning was performed with Bayesian optimization using Gaussian Processes, implemented on scikit-optimize [Head et al., 2021], with 100 calls to the optimizer. Further details on training hyperparameters and network architecture can be found in Appendix E.

7.5.2 Results

7.5.2.A Classification using Support Vector Machines and Logistic Regression

The results across the full range of configurations, including the different ASR, deviation-scores, correlation thresholds, and normalization strategies, are presented in Appendix E. The highest classification performance on both the development and test sets is reported in Table 7.9.

For PD detection, the highest classification performance on the development and test sets was achieved using deviation score DS_{RI}, without dimensionality reduction, with MinMax normalization, and a logistic regression classifier. This configuration attained 71.7% accuracy on the test set. If the 12 files excluded during data preprocessing (arbitrarily labeled as controls for a fair comparison with other works reporting on the entire dataset) were not included, the performance would increase to 73%. Each subject uttered 3 sustained vowels, so we also evaluated performance at the speaker level, after computing the majority vote of the three predictions per subject, resulting in 75% accuracy. The highest speaker-level accuracy was achieved by combining the deviation score $DS_{Q1,2,3}$, and the reduced dimensionality feature set with $CT = 0.9$.

For AD detection in ADReSS, the best classification results on the development folds were obtained using directly the features that constitute the full dimensionality feature set ($CT = 1$), combined with MinMax normalization, and an SVM classifier with second degree polynomial kernel, based on whisper transcriptions. This configuration reached 76% accuracy on the development folds, and 69% on the held-out test set. The best results on the held-out test set were obtained using the deviation score DS_{RI}, and reduced dimensionality feature set with $CT = 0.7$. Pre-processing and feature extraction using whisper transcriptions did not lead to the exclusion of any files from analysis. ADReSS only contains one picture

Table 7.10: Ablation study on the different variables for each configuration used in the disease detection experiments, using SVM and Logistic regression. Accuracy is reported in [%]. *test-wo-miss* refers to the test set disregarding the files that were excluded during pre-processing or feature extraction. *Features RN* refers to the features with "regular" normalization, instead of stratified by gender.

	Parkinson's Disease				Alzheimer's Disease		
	dev	**test-wo-miss**	**test**	**MV**	**dev**	**test-wo-miss**	**test**
ASR							
whisper	–	–	–	–	**62.7**	61.6	**61.6**
wav2vec	–	–	–	–	58.6	**63.1**	58.1
DS							
DS_{MSTD}	61.1	61.2	60.4	63.2	62.0	63.3	60.7
$DS_{MSTD-no-cap}$	60.1	60.1	59.4	61.9	61.5	66.1	63.4
DS_{Q123}	**64.8**	**65.1**	**64.2**	**67.4**	60.7	66.7	63.7
DS_{RI}	60.8	61.2	60.4	62.6	60.7	**66.9**	**64.1**
$DS_{Mahalanobis}$	60.2	59.4	58.7	61.5	54.1	58.8	56.6
Features	63.8	64.9	64.0	66.6	62.5	65.4	62.6
Features RN	63.6	64.5	63.6	66.0	**62.9**	49.3	47.8
CT							
$CT = 0.5$					**61.4**	**63.6**	**61.0**
$CT = 0.6$	61.3	61.3	60.6	63.2			
$CT = 0.7$					60.4	62.5	59.9
$CT = 0.8$	62.7	63.3	62.4	65.0	60.6	61.7	59.2
$CT = 0.9$	62.7	62.9	62.0	64.6	60.3	61.7	59.3
$CT = 1$	**63.1**	**63.9**	**63.0**	**65.7**	59.7	61.0	58.6
Normalization							
0-mean 1-var	62.2	62.8	61.9	64.2	60.6	61.3	58.8
MinMax	**65.4**	**65.4**	**64.5**	**68.1**	60.3	**63.3**	**60.7**
None	58.5	58.8	58.1	60.2	**61.0**	62.4	60.0
Classifier							
SVM	61.7	61.4	60.6	63.2	**61.9**	61.6	59.1
LR	**62.4**	**63.2**	**62.4**	**65.1**	59.3	**63.1**	**60.5**

description per subject, thus the performances reported at sample level are the same as at the speaker level.

Given the numerous variables involved in these classification experiments, Table 7.10 summarizes the average performance across all experiments for each ASR model, for each deviation-score, for each feature selection correlation threshold, for each normalization strategy, and for each classifier. Overall, whisper transcriptions yield better results than wav2vec transcriptions. This difference is partly due to wav2vec failing to generate a transcription for six files, which were excluded from further analysis. For consistency with other studies, the three files in the test set were treated as if the prediction was control.

In terms of deviation-scores, the best average performance for PD detection was achieved by $DS_{Q1,2,3}$. This trend is not observed for AD. One partial explanation is that, with whisper transcriptions, it was not possible to compute the $DS_{Q1,2,3}$ for the *discourse marker rate* feature. This issue arose because the bulk of discourse marker rate's data within the reference population corresponds to a very narrow range, leading to identical first and third quartiles, and resulting in an indeterminate $DS_{Q1,2,3}$. This feature is important for AD detection, as discussed earlier on this chapter, and its absence may hinder the performance of this score. Using the features directly as input to the classifier, i.e. without pre-computing a deviation-score, combined with regular normalization (instead of stratified regularization) yielded the

Table 7.11: Best disease classification results, using NAMs, in terms of accuracy (Acc), macro precision (P), macro recall (R), and macro F1, in [%].

				Dev Folds				Test				Test – Speaker MV			
CT	**DS**	**norm**	**ASR**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**
Parkinson's Disease															
CT=1.0	DS_{RI}	MinMax	–	75.0	75.8	74.9	74.8	68.7	69.9	68.7	68.2	73.0	74.7	73.0	72.5
Alzheimer's Disease															
C=0.5	Feats	MinMax	whisper	84.3	84.4	84.3	84.2	75.0	75.0	75.0	75.0	–	–	–	–

best average results on the ADReSS development folds. However, this approach fails to generalize to the held-out test set. The best results on the ADReSS held-out test set were obtained with DS_{RI}.

Regarding the decision to reduce the dimensionality of the feature set based on Pearson correlation between features, it appears that the detection of PD benefits from using the entire feature set, while the detection of AD befits from the reduced dimensionality feature set, with $CT = 0.5$. Notably, the dimension of the entire feature set used to study sustained vowels (20 features) is very similar to the dimension of the reduced set used for studying the picture description task (23 features).

The normalization strategy that appears to provide the best average results, for both PD and AD, is MinMax scaling, with the exception of the development folds in AD detection. Future work should further investigate the different normalization strategies, understand their impact, and discuss their inherent assumptions.

Finally, logistic regression appears to achieve better results than support vector machines. It is possible that the deviation-scores already provide substantial information, and thus a simpler classifier is sufficient. In fact, for AD detection, SVM achieves better results than LR on the development folds, but these are not generalized to the test set.

7.5.2.B Towards interpretable classification: results using NAMs

NAM experiments require more computational resources, thus only the configurations that yielded the best results in the previous section (both single best model, and best on average) were used to conduct the NAM experiments. The results on the complete set of experiments are reported in Appendix E. The best results are reported in Table 7.11.

For PD classification, NAMs were able to achieve 75% accuracy on the development folds, 69% of the test folds, and a speaker-level accuracy after majority vote of 73%. For AD detection, NAMs yielded 84% on the development folds, and 75% on the held-out test set. The NAM performance for PD detection on the test folds is lower than that achieved with LR/SVMs classifiers. Conversely, for AD detection, the NAM results are better that those obtained with LR/SVM classifiers (based on the best development performance).

More importantly than surpassing the classification performance, the advantage of NAMs lies in their inherent interpretability. The modularity and transparency of NAMs allow for precise visualization of what

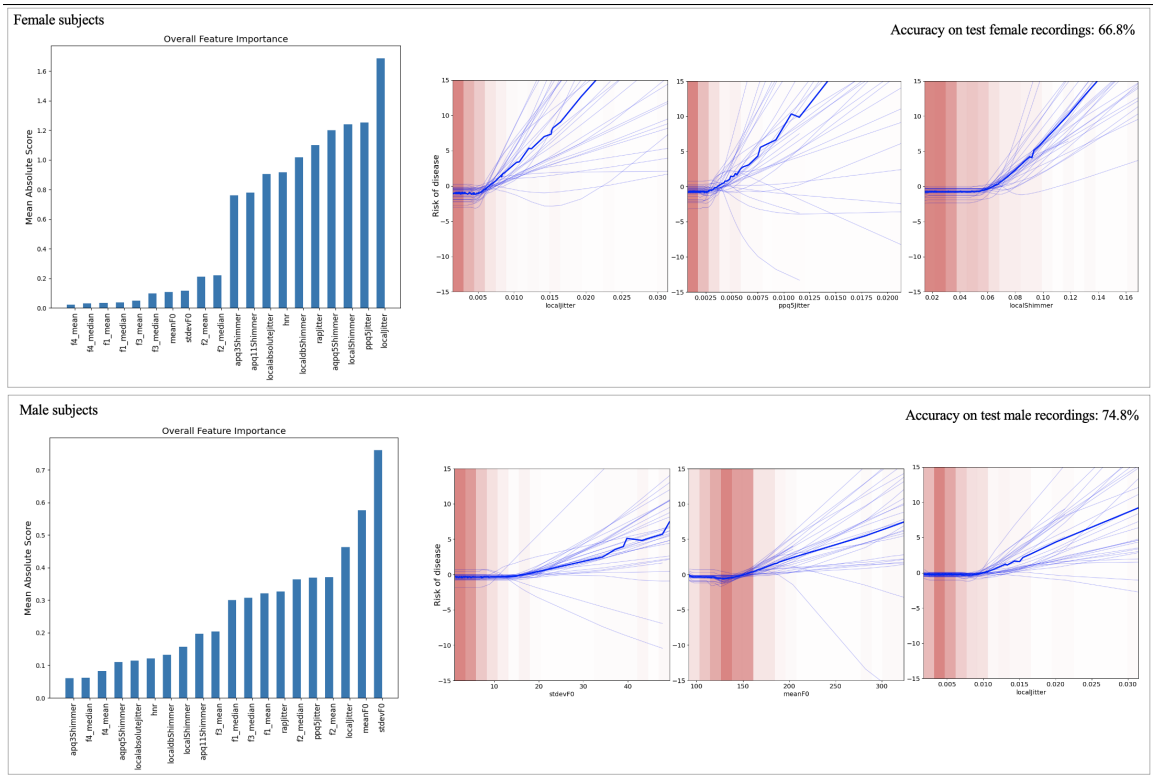


Figure 7.9: NAM trained for PD classification, for female (top) and male (bottom) subjects. The left plots represent the features that most contribute to the predictions; the plots on the right depict the shape functions learnt by the NAM, for the top three most important features.

the models learn during training and how each prediction is computed. Figures 7.9 and 7.10 depict the features that contribute most to the predictions, and present the corresponding learned shape functions. These graphs are not *a posteriori* explanations, but rather an exact description of how the model comes to a prediction. Each semi-transparent blue line corresponds to one model of the ensemble, trained on a given run of the cross-validation, i.e. the plots show 30 models (an ensemble of 3 models was trained for each of the 10 runs in the 10-fold cross-validation, as detailed in Appendix E). Following [Agarwal et al., 2021], the average score of each feature (each shape function) was set to zero, by subtracting the mean score, averaged over the entire training set. This results that, on binary classification tasks, positive scores increase the probability of the positive class, compared to the baseline probability of observing that class, while negative scores decrease the probability. On the same plots, the normalized data density is also visible, in the form of pink bars. The darker the shade of pink, the more data there is in that region.

One can observe substantial variability in the shape functions learned by each model, particularly when less data is available for training–indicated by lighter shades of pink–, and in the ADReSS corpus, which is half the size of PC-GITA. This variability gives a sense of confidence on the patterns learnt by, emphasizing the need for research with larger corpora to enable more robust conclusions.

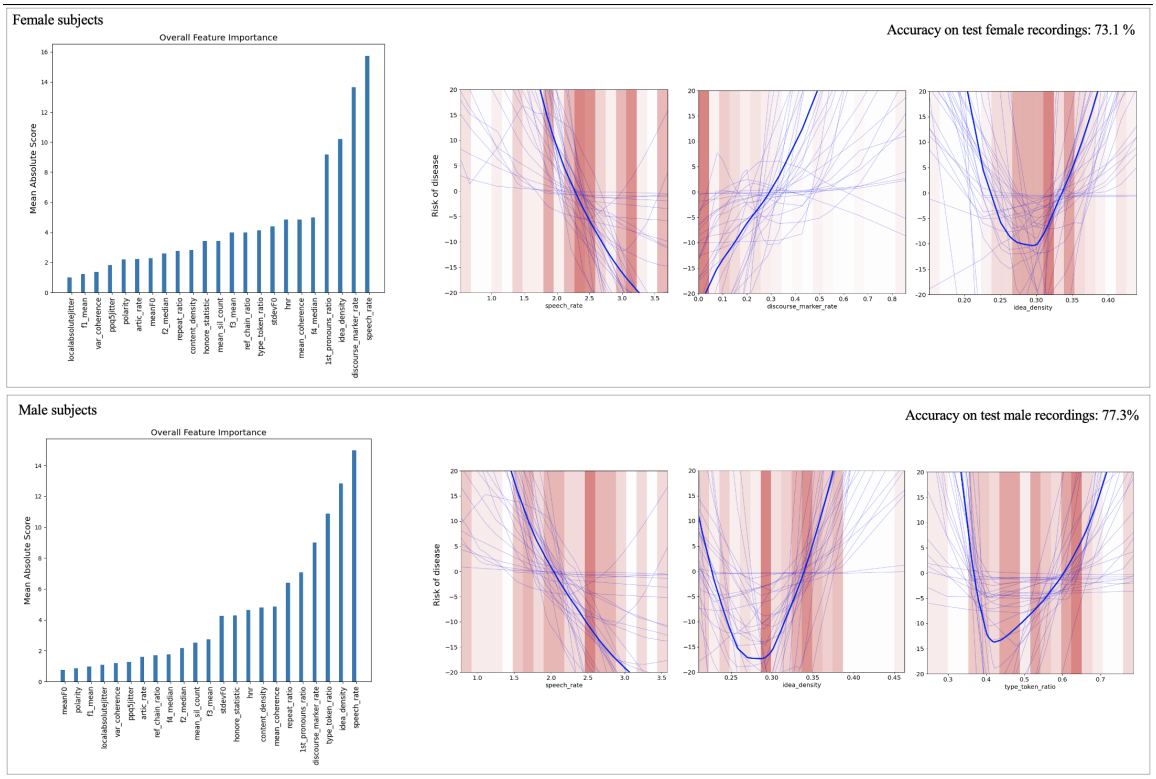


Figure 7.10: NAM trained for AD classification, for female (top) and male (bottom) subjects. The left plots represent the features that most contribute to the predictions; the plots on the right depict the shape functions learnt by the NAM, for the top three most important features.

Nevertheless, in most cases, the models learned curves that align with the expected manifestations of each disease in the speech signal

Upon analysing the NAM trained for PD detection which achieved the best accuracy on the test set, illustrated in Figure 7.9, it becomes evident that the outcomes differ between genders, although *local jitter* appears on the top three most important features for both genders. This feature is, by far, the most important for female speakers. The following features in terms of importance are also jitter and shimmer-related. This is consistent with expectations, as a high jitter and shimmer reflect cycle-to-cycle perturbations of F0, associated with impaired laryngeal control typical of PD patients. Previous studies have also found higher jitter values in PD patients when compared to control subjects (e.g. [Jiménez-Jiménez et al., 1997; Upadhya et al., 2017]).

For male subjects, the most important features are the standard deviation and mean of F0. The corresponding shape functions indicate that higher values of mean F0 and F0 standard deviation are associated with a higher risk of PD. Other works have also found increased F0 standard deviation in sustained vowels produced by PD patients compared to control subjects (e.g. [Goberman et al., 2002; Midi et al., 2008]). Goberman et al. [2002] suggested that this increase may be due to laryngeal instability, potentially caused by weakness of the laryngeal musculature resulting from rigidity or tremor.

The author mentions that tremor-related weakness has been found in other body systems, such as wrists [Brown et al., 1997]. Although not studied here, it is important to note that, in continuous speech, as opposed to sustained vowels, the F0 standard deviation is expected to decrease in PD patients, which is associated with the mono-pitch perception [Bowen et al., 2013; Harel et al., 2004; Ma et al., 2020] (see description in chapter 2).

The mechanism underlying the increased mean F0 in PD patients is suggested to be the increased rigidity of the laryngeal musculature (e.g. cricothyroid and thyroarytenoid muscles) [Duffy, 1995; Ma et al., 2020]. Biomechanical models of phonation demonstrate that increased vocal fold stiffness leads to higher fundamental frequency and jitter [Ma et al., 2020]. Various works have identified differences in mean F0 in PD patients and controls, although not always significant for both genders, nor in the same direction. For example, Goberman et al. [2002] found that mean F0 was higher in PD patients than controls, particularly in male speakers; Midi et al. [2008] found that mean F0 was higher in PD patients than healthy controls, but this difference was only significant in female subjects; and Yang et al. [2020] found the opposite, i.e., that mean F0 was lower in patients suffering from PD than controls. It is important to acknowledge that F0 is more than just a marker for vocal fold behaviour, it carries information about different speaker states and traits [Cummins et al., 2015b], and even physiological aspects such as hormonal balance and aging [Singh, 2019]. Thus, results should be interpreted with caution, and further research should be conducted with a larger dataset of age- and gender-matched controls and patients of PD and other diseases.

The patterns learnt by the NAM trained for PD detection are consistent with those represented on the radar plot of the sustained vowel for PD patients (Figure 7.5). Both the NAM and radar plot flag jitter and F0 as important features for characterizing the speech of female and male PD patients.

Upon examination of the NAM trained for Alzheimer's disease detection, which achieved the best performance, as depicted in Figure 7.10, it is evident that each model in the ensemble/cross-validation run learnt different patterns, yet some general trends can be inferred when averaging the predictions of those models. The first consideration is that the foremost contributing feature to prediction, for both male and female subjects, is *speech rate*. It is clear that the slower the speech rate, the higher the risk of the person suffering from AD. This behaviour is expected, as slower speech with more pauses is expected to be associated with a higher risk of AD. In fact, other works (e.g. [Hecker et al., 2022; Hoffmann et al., 2010]) have also identified the importance of speech rate for AD detection from speech.

Idea density is also among the top three most important features for detecting AD in both female and male speakers. The shape function learnt reflects a U-shape, indicating that low and high values of the feature idea density are associated with a higher risk of Alzheimer's disease. Low idea density has been associated with Alzheimer's disease at least since the well-known "Nun Study" by Snowdon et al. [1996], which found that low idea density in early life strongly predicted reduced cognitive ability or the

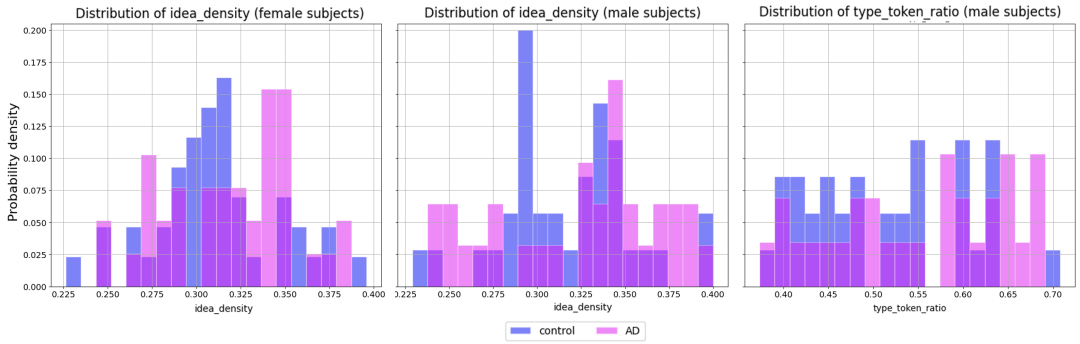


Figure 7.11: Distribution of idea density in male and female subjects, and TTR in male subjects, in ADReSS, for both controls (blue) and AD patients (pink).

presence of AD in later life. Boschi et al. [2017] conducted a comprehensive review and reported that AD patients have significantly lower idea density compared to controls.

A similar U-shape pattern was learnt by the subnet attending to the TTR feature, which is one of the top-three most important features for AD detection in male subjects. TTR, a measure of lexical diversity, is also associated with cognitive impairments. For example, Bucks et al. [2000] found TTR to be significantly lower in AD patients compared to control subjects. Berisha et al. [2015] suggested that measures of lexical diversity, including the number of unique words, are strong predictors of pre-clinical AD onset.

Contrarily, the fact that the NAM learnt to associate high values of these features with a higher risk of AD is more surprising, and not frequently reported in the literature, to the best of our knowledge. However, this pattern observed by the NAM is present in the data, as depicted in Figure 7.11. We manually inspected data samples labelled with AD associated high idea density and/or TTR and identified several patterns: some examples, although not very frequent, included whisper hallucinations; some examples corresponded to correct transcriptions, but were confusing or nonsensical despite high idea density or TTR; and occasionally, they were perfectly coherent descriptions of the Cookie Theft picture. These findings reinforce the idea that idea density, or TTR alone are not sufficient to make a prediction.

Nevertheless, this illustrates the advantage of having a fully transparent model, despite its imperfections. For instance, let us consider a scenario where a new individual undergoes testing with this system and receives a prediction of Alzheimer's disease. A healthcare practitioner could examine the reasons provided by the system for this prediction. If the sole reason provided was a high idea density or high TTR, the healthcare practitioner would have the information needed to make an informed decision or recommend further testing. Such reasoning would not be possible with a black-box model, or with a model that operates on uninterpretable features.

Finally, *Discourse marker rate* is also among the top-three most influential features for AD detection in female subjects. As the discourse marker rate increases, the risk of Alzheimer's Disease also increases, consistent with the findings of [Boschi et al., 2017].

7.6 Limitations

We acknowledge that this exploratory work has some limitations, and we encourage further research to address these potential drawbacks. Some of the features used, particularly vocal tract and voice-quality features, have been noted for their limited robustness across diverse recording conditions, including various devices–especially mobile platforms–background noise, and reverberation [Dineley et al., 2023; Jannetts et al., 2019; Maryn et al., 2017]. Therefore, these features may not consistently yield reliable results across corpora recorded under different conditions. We advocate for the establishment of guidelines to standardize how researchers record corpora and extract these features, which would enhance robustness and facilitate fair comparisons among different studies.

Another limitation pertains to corpora availability. The CLAC dataset, while valuable and larger than most speech corpora in clinical research, is crowdsourced, resulting in noise and lack of medical verification despite data filtering. Additionally, its size conditions the reliability of results, particularly for RI estimation, which ideally requires a minimum of 400 subjects per gender and age range. The small size of PC-GITA and ADReSS also adversely affects disease detection, specifically impacting the shape functions learned by NAMs' subnets. It is noteworthy that different hyperparameters result in different feature contributions and shape functions, and research with a larger dataset is essential to enhance robustness of results.

A third limitation relates to data normalization. To facilitate meaningful comparisons of RIs across datasets, we applied zero-mean and unit-variance normalization. Ideally, under consistent recording conditions, uniform speech task instructions, and robust feature extraction methods, such analyses could proceed without shifting the underlying data distribution.

Future studies should also investigate alternative methods for deciding when to partition the reference population prior to reference interval estimation and consider additional factors, such as level of education, accent, race, or body mass index.

Finally, another limitation of this study is that AD and PD detection tasks were addressed separately due to differences in the publicly available datasets. These datasets vary in speech tasks and recording conditions, which substantially influence the features extracted [Botelho et al., 2022, 2023]. Future work could explore simultaneous detection using the proposed framework when datasets with comparable tasks and recording conditions become available.

7.7 Summary

This work introduced a framework for the use of speech as an interpretable biomarker for multiple diseases. Although this chapter focuses on Alzheimer's and Parkinson's diseases, it proposes ideas that are suitable for using speech as a biomarker in general, including the screening of other speech affecting

diseases or even general health perturbations not typically categorized as diseases. This work starts by discussing that speech affecting diseases should not be regarded individually for two reasons: (1) they are risk factors for each other and thus frequently co-exist, and (2) they often have overlapping effects on the speech signal. Therefore, it is argued here that a valuable first step is to characterize the speech of control subjects. This characterization is based on reference intervals, a concept common in clinical laboratory science, but novel in the field of speech analysis for disease detection and monitoring.

In this study, reference intervals were established for a reference population. Nevertheless, our vision encompasses the potential of individualized definition of reference speech. This self-definition would facilitate precise identification of early signs of disease, and would enable personalized healthcare.

The initial feature set was defined to capture manifestations of various speech-affecting diseases, focusing exclusively on interpretable features. Yet, a high-dimensional feature space hinders the interpretability of results. Therefore, feature selection was conducted based on a reference population rather than datasets used for disease detection, for two primary reasons. Firstly, the goal was to define a feature set that captures reference speech, before identifying deviations specific to individual diseases. Secondly, supervised feature selection on disease detection datasets can lead to unstable results, overfitting, and poor generalization, particularly with small datasets [Dernoncourt et al., 2014; Soares et al., 2016; Vabalas et al., 2019]. The proposed approach yielded a concise feature set that capture different dimensions of reference speech: 8 features to characterize recordings of sustained vowels, and 23 features to characterize recordings of a picture description, enhancing interpretability. However, during classification tasks, while the reduced feature set was preferred for analyzing picture description tasks, the full feature set was preferred for sustained vowels. Future research should expand the initial feature set to include additional knowledge-based features, capturing broader dimensions of reference speech, and further validate these features with other corpora for disease detection.

Finally, the definition of reference speech was leveraged for the detection of AD and PD, by comparing how much controls and patients deviate from the reference population. Although the classification performance, measured as accuracy on the test set, falls below other works in the literature, we advocate for the exploration of this approach due to its transparency, thereby advancing speech as a reliable biomarker for health. In fact, it is well-documented that small sample sizes in clinical speech analysis studies often lead to overoptimistic estimates of model performance [Berisha et al., 2022; Ozbolt et al., 2022]. Therefore, we underscore the importance of interpretable outcomes. Particularly, the shape-functions learnt by NAMs correspond exactly to the decision process, instead of *a posterior* explanations. This transparency is crucial not only as a "second opinion" for clinicians, but also for early-stage research into speech as a biomarker. It facilitates multidisciplinary discussions among teams regarding the validity of model assumptions, and informs decisions regarding subsequent iterations, including data collection and feature refinement. Moreover, NAMs are suitable for multitask learning, enabling simulta-

neous detection of multiple diseases provided there is annotated speech data across different diseases, for the same speech tasks.

8

The advent of LLMs: macro-descriptor extractors for disease detection

Contents

THIS chapter explores the potential of Large Language Models (LLMs) as annotators of high-level characteristics of speech transcripts, which may be relevant for detecting diseases that affect language, particularly Alzheimer's disease. These low-dimension interpretable features, here designated as macro-descriptors (e.g. text coherence, lexical diversity), are then used to train a binary classifier. Our experiments compared the extraction of these features from both manual and automatic transcriptions, and involved both open and closed source LLMs, with several prompting strategies. The experiments also compared the use of macro-descriptors with the direct prediction of Alzheimer's disease by the LLM. Even though LLMs are not trained for this task, our experiments show that they achieve up to 81% accuracy, surpassing the baseline of previous Alzheimer's disease detection challenges, particularly when used as extractors of macro-descriptors. Furthermore, we explored whether providing information about pauses together with the textual information can improve AD detection.

8.1 Introduction

Recently, LLMs have transformed the landscape of artificial intelligence research and even permeated into the general public. In fact, LLMs have exhibited impressive language generation and understanding capabilities, with remarkable performance in numerous tasks, providing an opportunity to reconsider the potential of interpretable machine learning. Indeed, their ability to provide answers in natural language allows them to enlarge the extent of explanations that can be provided to a human. The natural question that arises then is: *are LLMs able to detect if a person suffers from a speech affecting disease that manifests itself in language perturbations, simply by analysing the transcriptions of the person's speech?* This is the question that we investigate in this chapter, particularly for the detection of Alzheimer's disease. We explore three approaches. In the first approach, which can be regarded as a baseline task, we define a set of prompts and query different LLMs as to whether each subject suffers from AD. The LLMs used were not trained nor fine-tuned specifically for this task (available data would be too scarce), and thus their main expertise is not on disease detection. Conversely, they are particularly capable of generating text, and capturing its structure. Therefore, we propose a second indirect approach: instead of directly asking the LLM to predict whether a person suffers from AD, we use it to annotate high-level and low-dimension aspects of speech transcripts, which we designate as **macro-descriptors**. These can then be used as input features for a simple and interpretable binary classifier. The macro-descriptors are chosen to capture the subjective perception of language's properties that are relevant for AD detection, corresponding to higher-level characteristics of speech affected by AD, previously described in chapter 2, Figure 2.5. Finally, in a third approach we explore whether pause information can enrich the text transcripts used in the previous approaches, and improve AD detection. The motivation for this third set of experiments is two-fold: (1) given the high importance of speech-rate in NAM-based AD detection in

the previous chapter, we hypothesise that this feature can further complement the macro-descriptors annotated by the LLMs; and (2) Yuan et al. [2020] obtained the best results at the ADReSS challenge [Luz et al., 2020] by encoding pauses into speech transcriptions and fine tuning pre-trained language models, ERNIE and BERT, with these enriched transcriptions.

To the best of our knowledge, the most similar work to this one also exploits an LLM for fluency opinion extraction [Bang et al., 2024]. However, for the final AD prediction, a neural architecture that combines encoded representations of audio, text and opinions obtained from GPT models is proposed, thereby losing explainability in the decision-making process.

In summary, this chapter targets three research questions:

- Are LLMs already able to perform AD detection from speech transcrips?
- Can we harness the potential of LLMs to capture low-dimension macro-descriptors that describe and help differentiate between the speech of healthy subjects and subjects suffering from AD, providing an interpretable explanation?
- Can pause information further complement speech transcripts and aid LLM-based AD detection?

8.2 Method

We propose two tasks aligned with the above research questions. Task 1 can be considered as a baseline, and consists of directly querying the LLM whether each person suffers from AD. Task 2 consists in using the LLM as an annotator of macro-descriptors, which are used as input to a binary classifier. We further investigate whether including pause-related information can aid AD detection. We compare five open and closed source LLMs: *Llama-2-13B* [Touvron et al., 2023] (henceforth referred to as **Llama**), *Mistral-7B-Instruct-v0.2* [Jiang et al., 2023] (**Mistral**), *Mixtral-8x7B-Instruct-v0.1* [Jiang et al., 2024] (**Mixtral**), *GPT-3.5-Turbo* [Ouyang et al., 2022][1] (**GPT-3.5**), and *GPT-4-Turbo* [Achiam et al., 2023][2] (**GPT-4**). To ensure reproducible predictions, we selected Greedy Decoding for the open-access models, and set a fixed seed and a temperature of 0.3 for the GPT models.

Additionally, and following the ideas presented in the previous chapter on the importance of defining reference speech, we define reference intervals for the macro-descriptors using a reference population, and qualitatively compare those intervals to the values of the macro-descriptors obtained for patients with AD and controls.

[1] Accessed using OpenAI's API in late February 2024.
[2] Accessed using OpenAI's API in early June 2024.

8.2.1 Corpora

The ADReSS corpus [Luz et al., 2020], previously described in chapter 3, is used as a test bed for the AD detection experiments. Data consists of speech recordings describing the "Cookie Theft" image, and corresponding manual transcriptions. The "Cookie Theft" image can be described using seven concepts, previously listed in chapter 3. We divide data into training / test sets with 108/48 subjects, respecting the partitions proposed in [Luz et al., 2020].

CLAC, also previously described in chapter 3, is used as reference population for defining the reference intervals for the same macro-descriptors. Here only the task where subjects describe the "Cookie Theft" image is considered.

8.2.2 Pre-processing

Although the audios released in the challenge were acoustically enhanced, we opted for using the original version. Interviewers' interventions were excised from both audio files and manual transcripts. Only plain text was retained in the transcripts, excluding non-speech events such as pauses and laughter. In addition to manual transcription, automatic transcription was conducted to simulate a scenario without available manual transcripts. Following the ASR results obtained in the previous chapter, in section 7.4.1.A, we report results with transcripts generated by both whisper and wav2vec. Wav2vec failed to output a transcription for 6 files (3 in train, and 3 in test).

8.2.3 Task 1: LLMs as AD predictors

In this baseline task, we prompt the LLMs to predict whether subjects suffer from AD, based on their description of the "Cookie Theft" image, employing four distinct prompting strategies. Additionally, a fifth prompting strategy frames the problem as a fluency evaluation task, similar to prior work [Bang et al., 2024]. In all prompting strategies, we ask for (1) a step-by-step explanation to facilitate chain-of-thought reasoning [Wei et al., 2022]; (2) the prediction (YES/NO); and (3) the confidence of the prediction (high/low), which is then used to breakdown the results by the level of confidence of the LLM. The prompting strategies, transcribed in full in Appendix F, are the following:

P1.1. 0-shot prompting, where we simply ask the model to predict if the person suffers from AD.

P1.2. 0-shot prompting, where we first explain the characteristics of speech uttered by someone with AD.

P1.3. 0-shot prompting, where we first explain the characteristics of speech uttered by someone with AD, and explain the seven concepts in the "Cookie Theft" image.

P1.4. Few-shot prompting, where we give the same explanations as in P1.3. and two examples, one description uttered by a healthy control, and one uttered by a person suffering from AD.

P1.5. 0-shot prompting, in which we frame the LLM as a fluency expert evaluator. We explain that the identification should focus on the same characteristics as those mentioned in P1.2-P1.4, and explain the seven concepts in "Cookie Theft". We then ask the model to identify if there are issues with the fluency (YES/NO), and use that as a proxy for AD prediction. The rationale behind this choice of prompting strategy is two-fold: first, LLMs may have guards to avoid making medical diagnosis, and second, LLMs are most likely trained on more data pertaining to generic fluency evaluation (e.g. language learning).

8.2.4 Task 2: LLMs as extractors of macro-descriptors

In this task we leverage the potential of LLMs to annotate *macro-descriptors*, i.e., high-level characteristics of the speech discourse that help differentiate between healthy controls and subjects suffering from AD. We define four macro-descriptors: **text coherence**, **lexical diversity**, **sentence length**, and **word finding difficulties**, which align with the higher-level manifestations of AD in speech, in Figure 2.5.

We define two prompting strategies for this task. In both, we explain the seven concepts in "Cookie Theft" and request, for each macro-descriptor, a real valued score between 0 and 1. The prompting strategies, transcribed in full in Appendix F, are the following:

P2.1. we frame the LLM as a fluency evaluator and ask it to rate the four macro-descriptors.

P2.2. we frame the LLM as a fluency evaluator in the medical domain, to support medical diagnosis. Besides the annotation of the macro-descriptors, we also request an AD prediction, and the confidence on the prediction. However, in this chapter, we only report the confidence analysis for task 1.

We evaluate the macro-descriptors qualitatively by comparing their distribution in both classes. Afterwards, we feed the four macro-descriptors as input features to a classifier, which aims at distinguishing healthy controls and AD patients. Following [Luz et al., 2020], we compare five binary classifiers: Support Vector Machines (SVM, with linear kernel and $C = 0.1$), Linear Discriminat Analysis (LDA), 1-Nearest Neighbour (1NN), Decision Trees (DT, with leaf size of 20) and Random Forest (RF, with 50 trees and a leaf size of 20). The hyperparameters were the same as described in [Luz et al., 2020] when specified, otherwise, default values were used. The classifiers were implemented in Python, using the scikit-learn toolkit [Pedregosa et al., 2011].

We trained the classifiers using 10-fold cross validation on the training set. Folds were defined to maintain a balanced distribution of classes and genders. Data were normalized to zero-mean and unit-variance, using the statistics computed on the 9 training folds. The predictions on the test set of the 10 classifiers trained during cross-validation were aggregated via majority voting, before computing the performance metrics.

Table 8.1: Example of manual transcription, enriched with pause information.

Transcription with pauses encoded
well (medium pause) this (short pause) uh (short pause) little (short pause) boy is (short pause) up (short pause) on (short pause) the stool (short pause) taking (short pause) cookies (short pause) handing (short pause) them down to his (short pause) sister and (short pause) she's (short pause) telling him to (short pause) be quiet (short pause) and (short pause) the stool (short pause) is (short pause) tipping (short pause) over (short pause) . the (medium pause) mother is (short pause) washing (short pause) or (short pause) drying dishes . the (medium pause) water is (short pause) running into (short pause) the sink (short pause) and (short pause) running (short pause) over (short pause) down (short pause) onto (short pause) the floor (short pause) . uh the (medium pause) wind must be (short pause) blowing because the curtains look like they're .. kitchen (long pause) door the (short pause) cabinet (short pause) door is (short pause) open (short pause) . mother's (medium pause) standing in water (short pause) . that's it (short pause) .

8.2.5 Introducing pause information

The results discussed in the previous chapter have highlighted the importance of *speech rate* information for AD detection. Other works have discussed that pause related information is relevant for studying cognitive impairment through speech. Particularly, Yuan et al. [2020] achieved the highest classification performance on the ADReSS challenge [Luz et al., 2020], by enriching transcriptions with pause information. Thus, we repeat task 1 and task 2 using transcriptions enriched with the information capturing pause duration. To this end, we computed words' timestamps by performing the forced alignment of manual and whisper-generated transcriptions with the corresponding speech signals. We employed the wav2vec pre-trained model from the *torchaudio* library[3] based on Connectionist Temporal Classification (CTC) segmentation, since a preliminary inspection revealed that word timestamps generated by whisper were not accurate enough. This observation was also encountered in the literature [Yang et al., 2024].

Following the approach of Yuan et al. [2020], pauses were grouped into three bins: short (under 0.5 s), medium (0.5-2 s), and long (over 2 s), with pauses under 50 ms excluded. In [Yuan et al., 2020], all punctuation was removed, and the three bins of pauses were represented by three punctuation marks–",", ".", and "..."–for short, medium, and long pauses respectively. In our experiments, we opted to code the three bins of pauses by explicitly writing "(short pause)", "(medium pause)", or "(long pause)" in the transcripts, as illustrated in Table 8.1. We believe this approach is clearer for prompting a LLM, than the punctuation-based encoding. Early experiments with alternative strategies to encode pause information (e.g. using the duration of the pause, or encoding only medium and long pauses) failed to outperform this strategy.

Additionally, we repeated the classification experiments from task 2 using as input features the four macro-descriptors combined with speech rate, computed with *Praat*, as in chapter 7. We opted to combine the macro-descriptors solely with speech rate, rather than other rhythm-related features, for two reasons. First, speech rate was identified as the most informative feature for Alzheimer's disease detection when utilizing neural additive models, as discussed in the previous chapter. Second, maintaining low dimensionality is crucial for interpretability.

[3] https://pytorch.org/audio/stable/tutorials/forced_alignment_tutorial.html

8.3 Results

8.3.1 Are LLMs adequate for AD Detection?

Table 8.2 reports the performance of the LLMs at predicting AD, for the train and test set. Results obtained with LLama-13b exhibited suboptimal performance and are consequently relegated to Appendix F. The results in the table are based on inference only, and the word “train” refers merely to the subset name. The table details the number of failed predictions (*#Fail*) by each model, i.e., instances for which the LLM did not output an AD prediction in the required format. For these instances, the label ”control” was arbitrarily assigned, reflecting a 50% a priori probability of a correct prediction due to the balanced nature of the dataset. This approach was chosen, instead of simply excluding these files from analysis, to maintain consistency across experiments by ensuring all analyses were conducted on the same data.

Examples of failed predictions range from outputting words such as ’undecided’ or ’maybe’ (6.96% / 5.23% of total requests for Mistral and Mixtral, respectively), to outputting hallucinations (8.98% / 1.26%). Most hallucinations originate from P1.4 with Mistral, where the model struggles to conform with the requested format. In fact, in this scenario, the model was only able to make a prediction for solely half of the instances. However, if failed predictions were discarded, P1.4 would achieve the highest classification accuracy: 78% on train, using Mistral and wav2vec transcriptions; and 85% on test, using Mistral and manual transcriptions. Conversely, when using the prompting strategy P1.5 with Mistral and Mixtral, we obtain less failed predictions than with P1.1 to P1.4, which may indicate that indeed the models struggle with presenting outputs for medical diagnosis. P2.2 also fails less frequently, possibly because the format is more strict, without giving the LLM the space to make open comments. In contrast, GPT-3.5 and GPT-4 were always able to follow the instruction. In the best case scenario, GPT-4 achieves an accuracy of 77% on the test set (using P1.4 on whisper transcriptions) and 75% on the train set (using P2.2 on manual transcriptions), with 0 *#Fails*.

On average, we observe that the larger the model, the less *#Fails* it makes. Mistral 7B, Mixtral 8x7B, and GPT-3.5 achieve similar average accuracy, while GPT-4 clearly outperforms the remaining models in terms of average accuracy. Overall, the table also shows some inconsistencies between the results on the train and test sets.

The prompts also request the confidence of the LLM in the prediction. We break down the accuracy of the AD prediction per confidence level in Figure 8.1. We note that, although we asked the LLM to output the confidence level using the words “high” or “low”, it often outputted other words, the most frequent being “medium”. In the figure, all non *high* levels were grouped as *low*. Our analysis suggests that, unlike GPT-3.5, which appears mostly overconfident in its predictions, Mistral, Mixtral and GPT-4 demonstrate an ability to gauge their confidence levels, in most cases. This implies the potential for

Table 8.2: AD classification based on LLM predictions. *#Fail* denotes the number of examples for which the model failed to follow the output instruction (identified on train/test). Reported results for accuracy, *Acc*, are presented in %.

	Mistral 7B			Mixtral 8x7B			GPT-3.5			GPT-4			Mean
	#Fail	Acc_{train}	Acc_{test}	#Fail	Acc_{train}	Acc_{test}	#Fail	Acc_{train}	Acc_{test}	#Fail	Acc_{train}	Acc_{test}	**Acc**
Manual transcriptions													
P1.1	9/4	61.1	64.6	10/5	55.6	54.2	0/0	65.7	54.2	0/0	67.6	70.8	62.0
P1.2	21/4	63.9	64.6	4/2	62.0	64.6	0/0	51.9	60.4	0/0	66.7	70.8	62.3
P1.3	15/6	61.1	75.0	8/3	72.2	60.4	0/0	54.6	56.3	0/0	70.4	70.8	64.9
P1.4	56/28	62.0	70.8	17/7	61.1	66.7	0/0	59.3	66.7	0/0	63.0	72.9	63.8
P1.5	2/2	55.6	60.4	0/0	53.7	50.0	0/0	50.0	50.0	0/0	50.0	52.1	52.6
P2.2	1/0	54.6	54.2	0/0	60.2	54.2	0/0	67.6	70.8	0/0	**75.0**	64.6	63.3
Whisper transcriptions													
P1.1	13/4	67.6	58.3	14/5	60.2	56.3	0/0	65.7	64.6	0/0	72.2	68.8	65.1
P1.2	17/3	64.8	70.8	5/7	63.0	54.2	0/0	65.7	72.9	0/0	70.4	75.0	66.7
P1.3	14/9	67.6	60.4	11/4	60.2	68.8	0/0	63.9	66.7	0/0	65.7	72.9	65.2
P1.4	54/21	61.1	70.8	11/4	66.7	66.7	0/0	63.9	72.9	0/0	63.9	**77.1**	66.3
P1.5	6/1	62.0	72.9	1/1	55.6	58.3	0/0	50.0	50.0	0/0	55.6	56.3	56.9
P2.2	3/2	55.6	52.1	0/0	62.0	56.3	0/0	68.5	75.0	0/0	71.3	70.8	64.1
Wav2vec transcriptions													
P1.1	21/7	58.3	52.1	18/8	56.5	52.1	0/0	54.6	52.1	0/0	60.2	52.1	55.8
P1.2	8/8	58.3	54.2	8/7	56.5	62.5	0/0	50.9	47.9	0/0	53.7	54.2	54.8
P1.3	16/6	56.5	54.2	10/3	61.1	58.3	0/0	50.9	47.9	0/0	57.4	58.3	55.9
P1.4	56/20	64.8	58.3	12/5	66.7	66.7	0/0	58.3	56.3	0/0	59.3	60.4	61.7
P1.5	0/0	49.1	45.8	0/0	51.9	47.9	0/0	50.9	47.9	0/0	50.9	47.9	49.7
P2.2	2/3	55.6	56.3	0/0	63.0	60.4	0/0	63.9	62.5	0/0	61.1	62.5	60.7
Mean	17/7	60.0	60.9	7/3	60.4	58.8	0/0	58.7	59.7	0/0	63.0	64.4	
Mean (man+wh)	18/7	61.4	64.6	7/3	61.0	59.2	0/0	60.6	63.4	0/0	66.0	68.6	

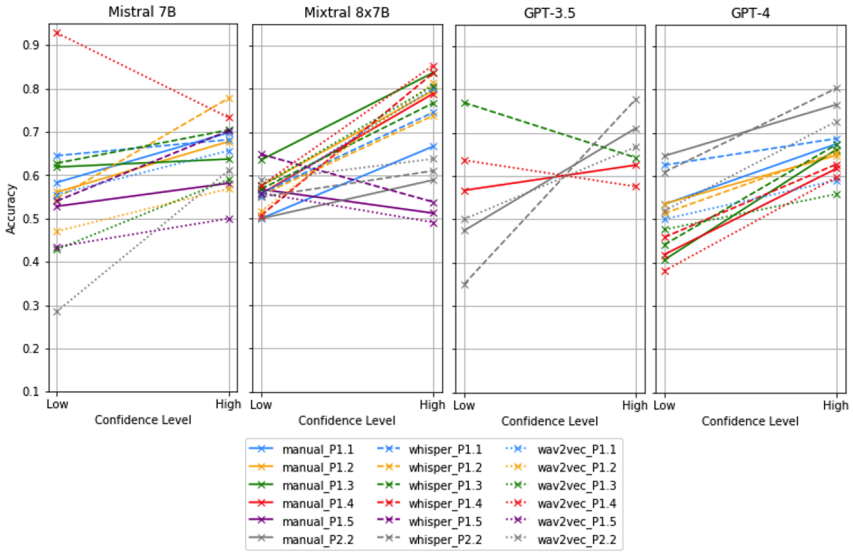


Figure 8.1: Task 1: Combined train and test accuracy per confidence level. A minimum of 10 instances per confidence level is required to include the prompting strategy in the figure.

leveraging the confidence level of the LLM when building models for disease detection. In fact, when considering only high-confidence instances, the best performing combination (wav2vec trancriptions fed into Mixtral, using prompt P1.4) reaches an overall accuracy of 85.4% (inference on train and test data together).

Furthermore, it is noteworthy that for Mixtral and GPT-4, prompt 1.5, which frames the LLM as a

Table 8.3: Prediction example, by Mistral 7B, based on prompting strategy P2.2, and Whisper transcripts.

Transcription
I don't see nothing but some roots. It's like somebody took some pencils or something and went up and down those things. Oh, I see a girl standing there or something. Some little knots or something on there. Oh, a lot of it around here. Some kind of little flower. And a sun. And a sun. And a girl is there. And there's something else over there. There's another girl. Look like... Look like some old girl is in there. I don't see nothing but some marks and things. Look to me about the same, except them things up there. I see this thing all look about the same to me, except this thing here. Look like a little kid or something. I'm sorry I didn't bring my glasses.
Coherence 0.3 **Word finding difficulties** 0.8 **Lexical Diversity** 0.5 **Sentence Length** 0.6 **AD Prediction:** YES

fluency evaluator outside the medical domain, results in the poorest performance in terms of confidence acknowledgment. Specifically, using prompt 1.5, Mixtral performed worst for high-confidence instances, and GPT-4 consistently exhibited overconfidence. This suggests that framing prompts within the medical domain may enhance the LLMs' ability to assess their confidence accurately.

8.3.2 LLMs as extractors of macro-descriptors

Task 2 explores whether LLMs can be used as extractors of macro-descriptors, which may be used as features for the detection of AD. Table 8.3 shows an example speech transcript and the corresponding macro-descriptors estimated with our best approach.

The distributions of the four macro-descriptors estimated with our best approach can be found in Figure 8.2. The figure suggests that the *Coherence* and *Lexical Diversity* macro-descriptors exhibit more pronounced differences between AD and control subjects, as evidenced by the distinct shifts in their distributions. However, for *Sentence Length* and *Word finding difficulties*, only AD patients display extreme values. Besides, the distributions in the figure appear to indicate that Mistral scores for the four macro-descriptors follow the trends expected from the current clinical literature.

Figure 8.3 shows the mean (dark green line) and reference intervals (shaded region delimited by lighter green lines) for each macro-descriptor, computed from the reference population. The figure also displays macro-descriptors estimated using the best approach for control subjects (blue) and subjects with AD (magenta). Unlike the approach in the previous chapter, the macro-descriptors were not normalized. Even without stratified normalization, subjects with AD deviate from the reference intervals more frequently than control subjects. This suggests that macro-descriptors estimated using LLMs may be robust to dataset shifts. This robustness may be due to macro-descriptors capturing very high-level discourse properties and the zero-shot prompting that mitigates adaptation to dataset-specific characteristics.

Additionally, we further investigate the validity of the macro-descriptors by using them as input features for classification experiments. Table 8.4 presents the results obtained with the five classifiers used in the ADReSS baseline: SVM, 1NN, LDA, DT, and RF. As in task 1, results obtained with *Llama-13*

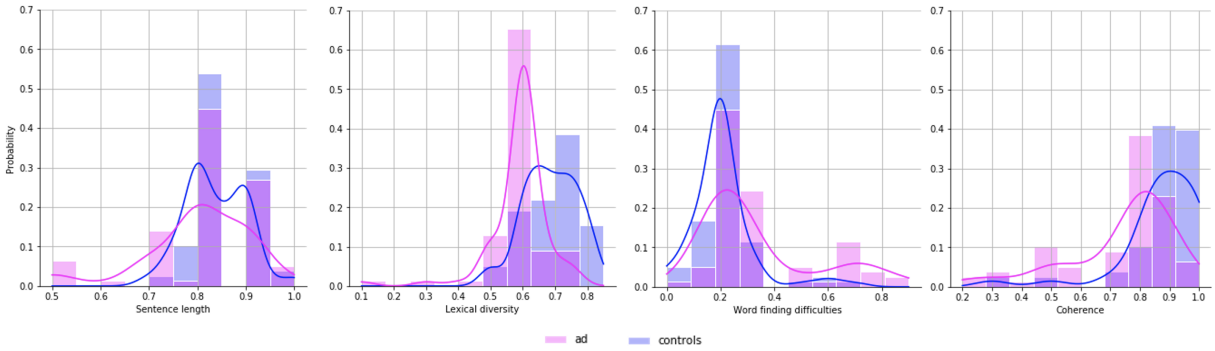


Figure 8.2: Distributions of the macro-descriptors, annotated by *Mistral*, as a response to P2.2, using whisper transcriptions.

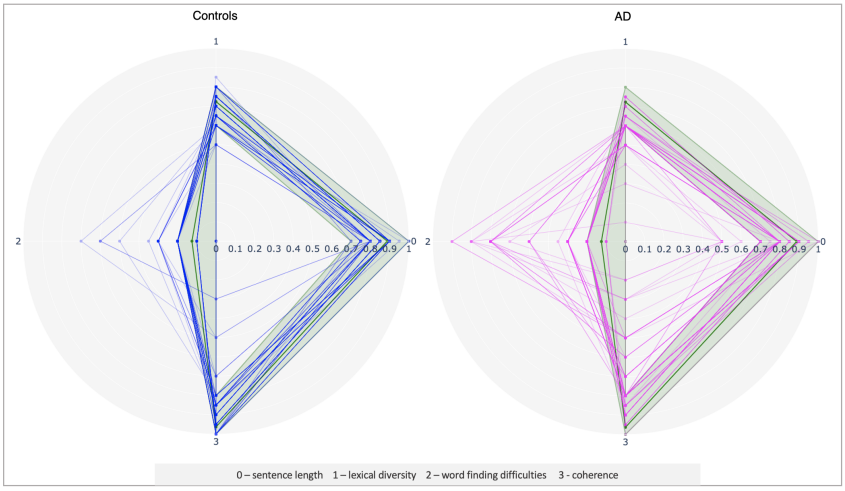


Figure 8.3: Radar plot that represents the reference interval for each macro-descriptor, in green, derived from CLAC. Blue lines correspond to the macro-descriptors values estimated for each control speaker, whereas magenta lines correspond to the macro-descriptors estimated for subjects suffering from AD.

were relegated to the Appendix F for brevity. The best performing combination in the 10-fold CV (using whisper transcriptions, P2.2, Mistral, and RF or DT) achieved an accuracy of 78.7%. On the test set, the best performance (whisper transcriptions, P2.2, Mistral or GPT-4, and SVM) reached an accuracy of 81.3%. These outcomes are directly comparable to other studies adhering to the train/test partitions of the ADReSS challenge, notably surpassing approaches based on features extracted with conventional methods, using the same classifiers and hyperparameters [Luz et al., 2020].

Overall, the table shows that the best performance was generally achieved using whisper transcriptions and prompt P2.2. Another general trend is that GPT-4 is, on average, the best performing LLM, followed by Mixtral, Mistral, and then GPT-3.5. Given the current concerns regarding data privacy, particularly in the healthcare domain, it is interesting to note that smaller and open-access models that can be run locally (Mistral and Mixtral) outperform GPT-3.5, which is larger and closed source.

Finally, Table 8.5 details the performance in terms of precision, recall and f-score for P2.1 and P2.2

Table 8.4: AD classification based on macro-descriptors.

		Mistral 7B					Mixtral 8x7B					GPT-3.5					GPT-4					**Mean**
		SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	
Train set: 10-Fold CV																						
Manual	P2.1	63.9	63.0	65.7	67.6	67.6	70.4	73.1	69.4	68.5	75.0	70.4	69.4	63.9	70.4	70.4	71.3	69.4	63.0	67.6	73.1	68.7
	P2.2	63.0	65.7	57.4	59.3	70.4	71.3	73.1	67.6	75.9	75.9	67.6	68.5	54.6	65.7	70.4	72.2	70.4	63.9	68.5	70.4	67.6
Whisper	P2.1	70.4	70.4	58.3	66.7	72.2	69.4	70.4	73.1	67.6	70.4	66.7	66.7	63.0	68.5	68.5	73.1	74.1	66.7	71.3	70.4	68.9
	P2.2	73.1	74.1	61.1	**78.7**	**78.7**	75.0	69.4	70.4	75.9	75.9	69.4	70.4	57.4	61.1	71.3	73.1	70.4	60.2	75.0	71.3	**70.6**
Wav2vec	P2.1	64.8	64.8	53.7	67.6	65.7	63.9	62.0	51.9	58.3	63.0	63.9	63.9	56.5	70.4	70.4	64.8	63.0	62.0	61.1	64.8	62.8
	P2.2	68.5	69.4	63.0	66.7	66.7	62.0	62.0	61.1	55.6	63.9	63.9	61.1	53.7	66.7	61.1	66.7	66.7	66.7	68.5	70.4	64.2
Mean				66.6					68.1					65.5					68.3			
Mean (man+wh)				67.4					71.9					66.7					69.8			
Test set																						
Manual	P2.1	70.8	66.7	45.8	68.8	68.8	66.7	77.1	66.7	64.6	79.2	58.3	60.4	58.3	66.7	62.5	75.0	77.1	60.4	72.9	77.1	67.2
	P2.2	68.8	60.4	54.2	64.6	70.8	77.1	75.0	70.8	75.0	75.0	68.8	68.8	60.4	70.8	68.8	72.9	72.9	77.1	68.8	72.9	69.7
Whisper	P2.1	68.8	68.8	43.8	72.9	72.9	64.6	66.7	64.6	72.9	72.9	66.7	66.7	70.8	68.8	68.8	79.2	79.2	70.8	79.2	79.2	69.9
	P2.2	**81.3**	77.1	64.6	**79.2**	**79.2**	70.8	62.5	64.6	75.0	75.0	75.0	72.9	66.7	77.1	72.9	**81.3**	79.2	66.7	75.0	77.1	**73.6**
Wav2vec	P2.1	62.5	62.5	70.8	62.5	66.7	72.9	68.8	60.4	70.8	77.1	70.8	70.8	62.5	66.7	66.7	66.7	66.7	58.3	64.6	77.1	67.3
	P2.2	62.5	62.5	60.4	62.5	68.8	64.6	72.9	58.3	68.8	72.9	62.5	62.5	54.2	58.3	62.5	64.6	68.8	52.1	68.8	66.7	63.8
Mean				66.3					70.1					66.3					71.6			
Mean (man+wh)				67.4					70.8					67.5					74.7			

Table 8.5: Precision, Recall and F1-Score for Whisper–Mistral–RF.

		class	**Precision**	**Recall**	**F1-Score**
P2.1	10F-CV	non-AD	72.2	72.2	72.2
		AD	72.2	72.2	72.2
	Test	non-AD	70.4	79.2	74.5
		AD	76.2	66.7	71.1
P2.2	10F-CV	non-AD	80.4	75.9	78.1
		AD	77.2	81.5	79.3
	Test	non-AD	81.8	75.0	78.3
		AD	76.9	83.3	80.0

using the best performing model on the 10-fold CV (whisper, Mistral and RF classifier). Results suggest that P2.2 is generally better than P2.1 in all dimensions. However, we can observe that the recall of the AD class tends to be higher than the precision of the AD class in P2.2, and the same is not true for P2.1. It may be the case that, when we frame the problem in the medical domain, the LLM behaves in such way that ultimately maximizes recall (or sensitivity, i.e., "Out of all the people that have the disease, how many got positive test results?").

Overall, using LLMs as extractors of macro-descriptors (Task 2) generally outperforms the direct AD classification of Task 1, on four dimensions: (1) by achieving higher accuracy on average, (2) by yielding smaller discrepancies between the results on the train and test sets, (3) by not failing to output predictions for the macro-descriptors, and (4) by providing a more interpretable approach, crucial in the healthcare domain.

Table 8.6: AD classification based on LLM predictions, from transcriptions enriched with **pause information**. *#Fail* denotes the number of examples for which the model failed to follow the output instruction (identified on train/test). Reported results for accuracy, *Acc*, are presented in %.

	Mistral 7B			**Mixtral 8x7B**			**GPT-3.5**			**GPT-4**			**Mean**
	#Fail	Acc_{train}	Acc_{test}	#Fail	Acc_{train}	Acc_{test}	#Fail	Acc_{train}	Acc_{test}	#Fail	Acc_{train}	Acc_{test}	**Acc**
Manual transcriptions													
P1.1	25/12	72.2	66.7	17/7	56.5	66.7	0/0	58.3	58.3	0/0	48.1	50.0	59.3
P1.2	11/5	55.6	70.8	10/1	65.7	68.8	0/0	56.5	54.2	0/0	50.0	52.1	58.3
P1.3	17/5	61.1	56.3	8/4	67.6	62.5	0/0	56.5	64.6	0/0	50.0	50.0	58.7
P1.4	52/20	67.6	68.8	5/0	71.3	58.3	1/0	73.1	70.8	0/0	53.7	54.2	65.4
P1.5	3/1	53.7	62.5	1/0	52.8	52.1	0/0	50.0	50.0	0/0	50.0	50.0	52.2
P2.2	0/0	52.8	54.2	0/0	67.6	68.8	0/0	63.9	70.8	0/0	50.9	50.0	59.5
Whisper transcriptions													
P1.1	19/6	63.0	60.4	5/8	58.3	52.1	0/0	55.6	62.5	0/0	50.9	52.1	56.9
P1.2	13/6	63.0	62.5	6/4	63.9	70.8	0/0	60.2	68.8	0/0	50.9	50.0	60.6
P1.3	21/9	64.8	66.7	4/6	63.9	66.7	0/0	60.2	68.8	0/0	50.9	50.0	60.9
P1.4	38/18	69.4	62.5	2/0	63.0	66.7	0/0	73.1	70.8	0/0	57.4	60.4	65.5
P1.5	3/0	53.7	62.5	1/1	54.6	52.1	0/0	50.9	50.0	0/0	50.0	50.0	52.7
P2.2	0/0	52.8	52.1	0/0	63.9	64.6	0/0	64.8	66.7	0/0	50.0	52.1	58.2
Mean	21/8	60.8	62.2	6/3	**62.4**	**62.5**	0/0	60.3	63.0	0/0	51.1	51.7	

Table 8.7: AD classification based on macro-descriptors extracted from transcripts enriched with **pause annotations**.

		Mistral 7B					**Mixtral 8x7B**					**GPT-3.5**					**GPT-4**					**Mean**
		SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	
Train set: 10-Fold CV																						
Manual	P2.1	63.9	68.5	55.6	65.7	68.5	70.4	65.7	68.5	71.3	66.7	66.7	66.7	57.4	69.4	67.6	66.7	75.0	56.5	64.8	67.6	66.2
	P2.2	67.6	71.3	55.6	67.6	70.4	70.4	70.4	66.7	70.4	64.8	67.6	63.9	59.3	50.9	62.0	67.6	74.1	66.7	69.4	68.5	66.2
Whisper	P2.1	69.4	69.4	60.2	73.1	72.2	72.2	72.2	62.0	69.4	66.7	69.4	70.4	68.5	75.9	75.9	69.4	73.1	63.9	68.5	69.4	69.6
	P2.2	71.3	68.5	59.3	73.1	71.3	69.4	69.4	65.7	63.9	68.5	72.2	71.3	61.1	75.0	75.0	72.2	73.1	66.7	75.0	72.2	69.7
Mean				67.1					68.2					**67.3**					69.0			
Test set																						
Manual	P2.1	62.5	66.7	41.7	62.5	66.7	66.7	66.7	68.8	66.7	77.1	60.4	60.4	66.7	70.8	70.8	79.2	68.8	47.9	64.6	83.3	65.9
	P2.2	62.5	66.7	56.3	66.7	66.7	72.9	72.9	72.9	72.9	70.8	77.1	77.1	72.9	72.9	79.2	70.8	68.8	66.7	64.6	75.0	70.3
Whisper	P2.1	68.8	68.8	47.9	68.8	75.0	72.9	72.9	47.9	72.9	75.0	62.5	64.6	52.1	68.8	68.8	72.9	72.9	66.7	72.9	75.0	67.4
	P2.2	79.2	81.3	50.0	81.3	81.3	72.9	68.8	68.8	70.8	72.9	68.8	75.0	58.3	70.8	72.9	70.8	72.9	64.6	66.7	68.8	70.8
Mean				66.0					70.2					**68.5**					69.7			

8.3.3 Can pause information complement LLM predictions?

Table 8.6 presents the results of task 1 using transcriptions enriched with pause information. Experiments were conducted using manual and Whisper transcriptions, as Whisper transcriptions consistently outperformed wav2vec transcriptions. Overall, this approach underperforms compared to using transcriptions without pause information (Table 8.2), except for the average performance of Mixtral. Surprisingly, despite its larger size, GPT-4 achieves the worst results.

Similarly, the results for task 2 with pause-enriched transcriptions (Table 8.7) are mostly inferior to those using original transcriptions (Table 8.5), with the exception of the average performance of GPT-3.

It is noteworthy that, unlike in [Yuan et al., 2020], the language models were not fine-tuned on pause-enriched transcriptions. The extra complexity of these transcriptions likely hinders meaningful information extraction, introducing distractions rather than value.

Table 8.8 presents the results for the alternative strategy to leverage pause-related information: com-

Table 8.8: AD classification based on macro-descriptors and acoustic feature **speech rate**.

		Mistral 7B					**Mixtral 8x7B**					**GPT-3.5**					**GPT-4**					
		SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	**Mean**
Train set: 10-Fold CV																						
Manual	P2.1	67.6	71.3	74.1	70.4	69.4	75.0	75.9	75.9	68.5	73.1	72.2	72.2	66.7	64.8	70.4	78.7	77.8	75.0	68.5	77.8	**72.3**
	P2.2	72.2	74.1	69.4	65.7	69.4	79.6	77.8	73.1	75.9	72.2	74.1	72.2	71.3	63.0	72.2	78.7	79.6	72.2	69.4	73.1	**72.8**
Whisper	P2.1	75.0	77.8	67.6	69.4	70.4	72.2	72.2	69.4	64.8	75.0	71.3	70.4	60.2	66.7	73.1	73.1	76.9	72.2	68.5	73.1	**71.0**
	P2.2	75.0	74.1	61.1	78.7	77.8	76.9	78.7	78.7	75.0	75.0	70.4	66.7	59.3	71.3	65.7	69.4	71.3	75.9	70.4	74.1	**72.3**
Mean				**71.5**					**74.3**					**68.7**					**73.8**			
Test set																						
Manual	P2.1	70.8	70.8	58.3	66.7	72.9	79.2	72.9	66.7	79.2	79.2	72.9	68.8	70.8	66.7	66.7	77.1	79.2	72.9	77.1	77.1	**72.3**
	P2.2	70.8	68.8	58.3	70.8	72.9	75.0	70.8	64.6	75.0	77.1	70.8	70.8	66.7	66.7	75.0	72.9	72.9	81.3	66.7	72.9	**71.0**
Whisper	P2.1	70.8	68.8	60.4	77.1	70.8	62.5	64.6	66.7	64.6	72.9	72.9	72.9	68.8	66.7	75.0	70.8	77.1	75.0	66.7	79.2	**70.2**
	P2.2	77.1	79.2	66.7	79.2	79.2	72.9	68.8	62.5	75.0	64.6	75.0	72.9	72.9	70.8	75.0	75.0	72.9	60.4	75.0	77.1	72.6
Mean				**70.5**					70.7					**70.9**					74.0			

bining *speech-rate* with the four original macro-descriptors from task 2 for training the binary classifiers for AD detection. Contrarily to the pause-enriched transcription approach, this method yielded higher cross-validation performance across all models, with Mistral and GPT-3.5 also achieving improved performance on the test set. The test performance decline for Mixtral and GPT-4 was minor, at 0.1% and 0.7%, respectively. These findings indicate that speech rate provides complementary information to the macro-descriptors, enhancing AD detection capabilities.

8.4 Summary

This work explores the potential of LLMs for AD classification, a task beyond their direct training scope. The experiments demonstrate that employing LLMs as extractors of macro-descriptors for AD (Task 2) compares favourably with the direct prediction of AD by the LLM (Task 1), in terms of average performance, failed predictions, and interpretability. Additionally, incorporating speech rate as an input feature alongside the four macro-descriptors in task 2 enhances AD classification.

Despite the encouraging results, a relevant limitation is the susceptibility of this approach to small prompt variations, resulting in divergent outputs, and influencing downstream performance. This lack of consistency aligns with observations across various tasks involving LLMs, motivating automatic prompt optimization strategies [Battle and Gollapudi, 2024]. Nevertheless, this type of approach is prone to overfitting, which may limit its application based on a small corpus such as ADReSS. Future research should thus involve the use of larger corpora, ideally from diverse settings. This research opens pathways for the use of LLMs in the development of interpretable approaches for disease screening, showcasing their significant potential in capturing discourse structure.

9

Conclusion

Contents

THIS thesis describes our exploratory work on the use of speech as a trustworthy biomarker to support accessible medical diagnosis. More than proposing a solution to a well-defined problem, this thesis reflects a journey towards understanding the multifaceted challenge of the automatic detection of speech affecting diseases.

This chapter begins with a summary of the principal findings and contributions of this thesis. Subsequently, we outline and discuss the most significant directions for future research. Considerations on ethics and privacy are included, emphasizing the necessity of regulations and ethical practices that researchers should adhere to. The chapter concludes with final remarks, including on the prioritization of reliability and explainability of results, and the risks and opportunities presented by new AI tools.

9.1 Summary of key findings

From the beginning, this thesis took a broad view of the exploration of speech as a biomarker. It adopted a holistic approach viewing speech as a biomarker across multiple diseases rather than focusing narrowly on one individual condition, and analyzing speech data collected in both controlled and naturalistic settings. It sought to identify specific speech characteristics affected by health conditions to aid healthcare providers. Additionally, the plan included complementing speech with signals from the brain and nervous system as potential biomarkers for neurological and cognitive disorders, aiming to deepen understanding of their intricate relationship.

Aligned with this last perspective, the starting point of this PhD concerned the exploration of EMG signals produced during speech articulation. In particular, we focused on a two-step approach to convert acoustic speech into the underlying EMG signal. We observed that our system was able to estimate with good performance four out of the five EMG time domain features, and that these features contained sufficient information to retrieve the underlying EMG signal. We further discussed the challenges associated with multi-session and multi-speaker experiments, including the variability induced by different electrode positions, and different skin, muscle, and fat properties. This work, together with [Diener et al., 2020], laid the ground for the emerging field of *Silent Computation Paralinguistics*. Other authors have continued these efforts, namely proposing alternative systems for the speech-to-EMG conversion [Scheck and Schultz, 2023; Ullah and Kim, 2024], and estimating prosody from EMG signals [Vojtech et al., 2022]. We anticipate that, when more data becomes available, the emerging field of Silent Computational Paralinguistics could contribute to disease detection. By exploring biosignals beyond acoustics, it may be possible to differentiate diseases based on which parts of the speech production process are most affected. In fact, the broader field of Silent Communications has seen substantial developments in the healthcare sector. A notable example is Unbabel Halo[1], which combines a non-invasive neural interface

[1] https://aiforgood.itu.int/speaker/unbabel-halo/

with generative artificial intelligence to transform biosignals into language, thereby helping patients with Amyotrophic Lateral Sclerosis stay connected.

Motivated by the COVID-19 pandemic, which restricted data collection for exploring the interconnections between biosignals in speech production and neurological or cognitive disorders, we redirected our focus to other remotely collectible biosignals highlighted as crucial during the lockdown. Specifically, we investigated facial images and visual speech as complementary modalities to acoustic speech for detecting OSA. Our previous research on Silent Computation Paralinguistics suggested the effectiveness of EMG signals in capturing articulation patterns for paralinguistics. Building on these findings, we hypothesized that embeddings trained for visual speech, or lip reading, would also contain paralinguistic information. These embeddings could encode articulation patterns, craniofacial structure, and breathing patterns, which are particularly informative for obstructive sleep apnea detection. Therefore, we introduced visual speech as a new modality for paralinguistics, achieving promising results.

Health-related speech datasets are scarce and difficult to acquire due to patient-privacy laws, ethical concerns, lack of awareness in the medical community, and resource constraints. The COVID-19 pandemic exacerbated these difficulties. Consequently, it became particularly relevant to explore large repositories of pre-existing data online. We thus leveraged vlogs with self-reported health status available on YouTube as an initial step towards multimodal disease detection, which could later be applied to telemedicine appointments.

Notwithstanding the promising results obtained with transfer learning and neural networks for in-the-wild multimodal obstructive sleep apnea detection, questions remain about what these models are actually learning and the basis for their predictions, which can significantly impact patients' lives. We argue that before systems such as the one we developed can transition to commercial applications, we need to ensure their reliability. Therefore, we conducted a series of experiments to foster the discussion on the challenges associated with datasets for disease detection. In particular, our experiments highlighted unexpected biases in datasets, and the difficulty of translating models and results to new domains, including different recording conditions, speech tasks, and languages. This difficulty gains an important dimension when we observe that the speech features typically used for disease detection encode substantial information about the recording conditions.

These challenges, along with the discussion on chapter 2 about multimorbidity and the overlapping effects of multiple diseases on speech, prompted us to work towards a framework for health monitoring that should be robust to dataset shifts, consider several diseases simultaneously to cope with multimorbidity frequent in an aging population, and provide explanations compatible with clinical reasoning. As an initial step, we characterized reference speech by defining reference intervals for clinically interpretable acoustic and linguistic features, drawing on a concept from clinical laboratory science. The CLAC corpus, used to define the reference population, is a valuable resource designed to characterize

English healthy speakers. However, it has limitations: it lacks sufficient speakers to provide distinct characterizations for each age range, and it is crowdsourced, resulting in diverse recording conditions and adherence to instructions. In our experiments, to compare this dataset with others for disease detection, we performed dataset-dependent normalization. We anticipate that future work with a larger corpus, including both healthy and diseased individuals across all age ranges, collected under controlled conditions, may provide a better characterization of reference speech without the need for dataset-dependent normalization. Furthermore, we propose a radar plot visualization tool to integrate all features. A single radar chart that describes the speech properties of each new subject and highlights deviations from the norm could be a valuable tool for supporting physicians in disease screenings.

Ideally, rather than defining reference speech for each subpopulation based on demographic criteria, it would be preferable to establish a self-definition of reference speech. Deviations from an individual's own reference speech would provide more sensitive indicators of early disease signs, thereby enabling truly personalized care.

Additionally, we explored Neural Additive Models, glass-box neural networks that elucidate the basis for model predictions. Despite other models achieving higher classification performance, NAMs offer the advantage of interpretability, providing meaningful clinical insights. This is crucial not only as a "second opinion" for a physician, but also in early-stage research on the use of speech as a biomarker, facilitating multidisciplinary team discussions on the model's assumptions, validity and trustworthiness, and informing decisions on further iterations, including data collection, or feature adjustments. NAMs are also suitable for multitask learning, enabling the simultaneous detection of multiple diseases, provided that data of the same speech tasks annotated across different diseases is available.

Our final set of experiments leverages the remarkable capabilities of large language models in text understanding, to annotate clinically meaningful high-level dimensions, termed macro-descriptors. Using just four macro-descriptors, we surpass previous results obtained with conventional features for Alzheimer's disease detection. These macro-descriptors are effective with simple models like SVMs, and could potentially enhance interpretability when utilized with neural additive models. Likely due to their high-level nature and the zero-shot prompting strategy employed for extraction, these macro-descriptors demonstrate increased robustness to dataset shifts and we hypothesize they may even offer resilience in multi-lingual scenarios. Moreover, by querying LLMs for the extraction of macro-descriptors that capture discourse structure instead of directly querying for a diagnosis, we mitigate the risks associated with biased or compromised advice from LLMs, as highlighted in recent reports concerning mental health applications [Ma et al., 2023]. Besides, the macro-descriptors can be complemented with acoustic information; in our experiments, speech rate improved Alzheimer's disease detection.

Overall, this thesis explored multiple perspectives on the potential of speech as a biomarker. We investigated novel modalities, such as EMG and visual speech, which complement the speech signal and

show promise for paralinguistics and disease detection. Additionally, we discussed several challenges associated with small datasets, emphasizing the need to focus on reliability and interpretability. Our work on reference speech characterization, and neural additive models for disease detection aimed at providing a holistic perspective on health, paving the way for future studies that could establish self-definitions of reference speech, and the simultaneous screening of multiple diseases. Finally, we leveraged the new LLMs to capture high-level discourse characteristics, bridging powerful black-box models and interpretable dimensions essential for clinical applications.

9.2 Future work

Technical challenges

The first priority of future work should be to establish a comprehensive protocol for health-related data collection, encompassing both speech and other non-invasive complementary modalities, and initiate data collection efforts. This protocol must specify rigorous yet practical recording conditions that can be implemented in typical facilities. It should standardize speech tasks relevant for various speech affecting diseases, and provide clear instructions for participants. Data should include annotations for multiple diseases, particularly those that are risk factors for each other, or have overlapping effects on speech, and represent all age ranges, genders, and biological sexes. It should also include annotations for possible confounding factors, such as education, and occupation, as well as symptoms, and risk factors. Although this is a substantial endeavour, it has the potential to revolutionize research on speech as a biomarker. In fact, the availability of very large datasets has been a key factor in enhancing the performance of numerous machine learning models, namely on speech and language tasks. Furthermore, existing publicly available datasets are not only limited in size and focused on single diseases, but also have become over-explored benchmarks, leading researchers to optimize for the benchmark rather than the problem they represent.

Recognizing the challenge of labeling large amounts of data for multiple diseases, one possible approach to manage the uncertainty of co-occurring diseases with high prevalence is to employ soft labels instead of one-hot encodings. Rather than assigning a zero probability of disease to a presumed control subject, soft labels can more accurately reflect disease prevalence within specific sub-populations based on demographics.

The second priority is to discuss, elaborate, and systematize the anticipated effects of diseases on the speech signal, along with the causal mechanisms responsible for such effects. This thesis has provided initial efforts in this direction, as illustrated with the diagram in chapter 2, Figure 2.5. However, we emphasize that this diagram represents hypotheses from the literature, and requires further validation by a multidisciplinary team, including (but not limited to) speech engineers and scientists, speech patholo-

gists, and physicians specializing in each of the speech affecting diseases. Afterwards, the hypotheses should be validated through data analyses.

One key aspect that becomes evident when analysing the diagram is that one possible approach for using speech as a biomarker for health is to focus on detecting certain symptoms, risk factors, and/or pathological mechanisms, rather than directly performing binary disease detection. These factors could, in turn, be used either for predicting the risk of disease, or as a parallel explanation for disease prediction. One approach could involve supervising the training of an intermediate layer in a neural network for disease detection, where each dimension represents a symptom, risk factor, or high-level manifestation of a disease, similar to the idea explored by Tu et al. [2017]. This layer could include dimensions such as overweight, aging, laryngeal trauma/inflammation, reduced lung capacity, formal thought disorder, etc.

Alternatively, if these annotations are not available, future work can use LLMs to capture high-level speech characteristics, or *macro-descriptors* suitable for multiple diseases, as we did for Alzheimer's disease. Our work with LLMs focused on text, but recent models like GPT-4o [OpenAI, 2024] are equipped to handle other modalities, including audio, making them promising for capturing both acoustic and linguistic macro-descriptors.

An important avenue for further research is causal machine learning, particularly in elucidating the causal mechanisms linking diseases with speech manifestations.

Although this thesis focuses mostly on binary disease detection, due to data availability, the assessment of disease severity as a regression task is equally important, particularly for applications in regular disease monitoring. The evaluation of severity, or progression of a disease, is further complicated by the scarcity of datasets that provide detailed annotations of disease severity and comprehensive representations of all disease levels.

Ethics and privacy challenges

Speech data inherently carries considerable personal information, as has been extensively discussed throughout this thesis. This personal information is exposed daily, without the speakers' control over which information is divulged and subsequently used by third parties possibly with unethical or malicious intentions. For instance, a scenario where an insurance company explores speech as a health biomarker to deny insurance coverage is clearly unethical [Singh, 2019]. Given the sensitive nature of speech data, although it is outside the scope of this thesis, it is important to mention some key ethics and privacy considerations essential for responsible research.

The entry of consumer tech companies into the health market, providing data collection, storage, and analysis solutions, raises issues relating to privacy, data protection and ownership, and informed consent [Cummins and Schuller, 2020]. The proliferation of such technologies blurs the distinction between medical and non-medical devices and raises further challenges, by complicating regulatory efforts [Cummins and Schuller, 2020].

Singh [2019] emphasizes the importance of the speaker's awareness and consent, when it comes to ethical uses of voice data. Information should not be distributed nor used without the speaker's explicit permission. Naturally, this requirement complicates data sharing across research institutions, potentially limiting dataset sizes and thus reducing model robustness and result reliability, thereby underscoring the importance of informed consent. Informed consent is particularly complex when subjects are children, psychiatric patients, or individuals with dementia.

Addressing these issues requires societal and legislative efforts, as the pervasive nature of speech complicates regulation [Singh, 2019]. Existing regulations, such as the General Data Protection Regulation (GDPR) [European Parliament and Council, 2016], and the AI Act [European Commission, 2021], offer a first step in this direction and should be adhered to.

Besides these issues that should be addressed by societies and lawmakers, researchers in the field of speech health should also consider ethical practices in their work. First of all, upon data collection, obtain informed consent from participants. Secondly, ensure that data is securely stored, and properly anonymized if data sharing is allowed. Finally, ensure ethical use of technology. Indeed machine learning applications are prone to learning biases in the data, and efforts should be made to mitigate these biases, ensuring that predictive models are transparent, fair, and prevent any form of discrimination.

Privacy-preserving voice processing algorithms are another relevant direction. These cryptographic-based technologies, which process voice data without exposing the signal itself, represent an emerging area of research [Singh, 2019].

A final ethical issue, distinct from the previous ones but equally deserving of philosophical consideration, is the extent to which individual patients benefit from early diagnosis or diagnosis at all. For instance, early diagnosis is certainly beneficial for a person suffering from a highly contagious respiratory disease with a high probability of recovery. However, is it beneficial for an elderly individual who may be diagnosed with dementia while still feeling well and "happy"? Does such a diagnosis bring value or sorrow to their life? We do not aim to provide an answer but rather to encourage debate among multidisciplinary teams and further reflection from our readers.

Final Remarks

Several findings of this thesis point to unreliable results in the literature for disease detection. This concern has also been recognized by a broader segment of the research community, which has documented issues such as overoptimistic results [Berisha et al., 2022; Espinoza-Cuadros et al., 2016; Ozbolt et al., 2022] and the Clever Hans effect [Liu et al., 2024]. This debate extends beyond speech-related health research to other high-stakes fields [Kapoor and Narayanan, 2023].

Nevertheless, rather than discouraging researchers, I believe this debate should mark a turning point towards responsible and trustworthy research for disease detection. Prioritizing reliability and explainability of results is crucial, even if at the cost of lower performance or simpler models. Many

studies describe results as *promising*. This should not be interpreted as a euphemism for failure but as a genuine indication of the potential for speech as a tool to improve healthcare access. Now is the time to direct this promise towards practical applications with the necessary caution. This field is at an exciting juncture, with tools of enormous potential ready to make a significant impact through interpretable and reliable models. The focus on explainability is vital. Kahneman [2011] introduced the framework of fast and slow thinking, where he distinguishes two styles of processing in our brain: *system 1* and *system 2*. System 1 corresponds to an unconscious process, which takes place very rapidly, and relates to implicit (intuitive) knowledge and the corresponding neural computations. System 2, on the other hand, corresponds to controlled and conscious processing, which involves a sequence of thoughts, which are usually verbalizable [Goyal and Bengio, 2020]. While some tasks can be achieved only with system 1's capabilities (e.g. riding a bike), others require the conscious controlled type of processing and the explicit knowledge of system 2 (e.g. imagination, planning, or discovering of causal dependencies). However, system 2's cognition generally also requires unconscious processing (system 1) to perform much of the work, sampling candidate solutions to a problem, from a possible astronomical number [Goyal and Bengio, 2020]). According to [Goyal and Bengio, 2020], current deep learning methods are fairly good at tasks associated with system 1 – "They can rapidly produce an answer [...] through a complex calculation which is difficult (or impossible) to dissect into the application of a few simple verbalizable operations. They require a lot of practise to learn and can become razor sharp good at the kinds of data they are trained on". Clinical experts' diagnosis process can also be explained in the light of system 1 and system 2 dichotomy. Brush Jr et al. [2017] describe that system 1 enables the quick retrieval of an exemplar stored in long-term memory to bring to mind diagnostic possibilities. This occurs automatically, naturally, and without conscious control. System 2 is used for testing, analyzing, and verifying a diagnostic hypothesis.

Likewise, I argue that designing a system for the automatic detection of speech-affecting diseases may benefit from a framework that incorporates both System 1 and System 2 capabilities. A non-interpretable large model, trained on extensive data or leveraging existing large language models, could identify disease candidates using correlations or risk factors (aligned with System 1 capabilities). For example, it would be acceptable to associate overweight middle-aged men with a higher risk of obstructive sleep apnea. Indeed there is a high incidence of OSA in overweight middle-aged men – and this is also similar to how a sleep doctor might reason to sample diagnostic candidates, when a middle-aged overweight man walks into their medical office.

Simultaneously, a parallel system aligned with System 2 should provide explanations compatible with clinical reasoning. This explanation does not need to provide the description of the exact process of the first system to reach a decision but should offer insights that guide further clinical decisions. For instance, it should explain why the individual is at high risk for OSA, considering not only demographics

or overweight risk, but also factors such as retrognathia, larynx inflammation, or impaired velum control. This approach enhances diagnosis understanding and informs subsequent medical recommendations. For the System 2 component, we can use neural additive models, LLM-annotated macro-descriptors, intermediate neural network representations capturing risk factors or high-level disease manifestations, and radar chart visualizations to highlight deviations from reference speech.

Before concluding this thesis, I invite the reader to further reflect on the dual nature of recent AI developments in healthcare, which present both substantial risks and remarkable opportunities beyond the detection of speech-affecting diseases. Among the risks are the potential for disseminating misinformation, providing harmful advice — as evidenced by instances where foundational models have encouraged engagement in eating disorder behaviors [Center for Countering Digital Hate, 2023] — and exacerbating healthcare access disparities. However, these risks can be mitigated through the implementation of appropriate regulations and enhancing public literacy. It is important that fear of these threats does not hinder the exploration of the considerable potential for social good offered by new AI tools. These tools have enormous potential, including enhancing health literacy for all, reaching remote populations, combating isolation, improving patient experiences, enabling personalized treatment plans [Coker et al., 2022], facilitating AI-enabled drug discovery [Center, 2023], and supporting the autonomy of individuals with dementia [Küster and Schultz, 2023].

Bibliography

A. Abad, A. Pompili, A. Costa, I. Trancoso, J. Fonseca, G. Leal, L. Farrajota, and I. P. Martins. Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech & Language*, 27 (6):1235–1248, 2013.

A. Ablimit, C. Botelho, A. Abad, T. Schultz, and I. Trancoso. Exploring dementia detection from speech: Cross corpus analysis. In *ICASSP*, pages 6472–6476. IEEE, 2022.

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

W. G. Admoni. *Der Umfang des Ganzsatzes und des Elementarsatzes im Deutschen*, pages 1–10. De Gruyter Mouton, 2019. doi: doi:10.1515/9783110850604-004. URL https://doi.org/10.1515/9783110850604-004.

R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton. Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems*, 34:4699–4711, 2021.

S. Al-Hameed, M. Benaissa, and H. Christensen. Simple and robust audio-based detection of biomarkers for Alzheimer's disease. In *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 32–36, 2016.

M. Al Ismail, S. Deshmukh, and R. Singh. Detection of COVID-19 through the analysis of vocal fold oscillations. In *ICASSP*, pages 1035–1039. IEEE, 2021.

American Press Institute. Readers' degree of understanding, 2009.

S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller. Snore sound classification using image-based deep spectrum features. 2017.

M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *Journal of neural engineering*, 16(3):036019, 2019.

G. K. Anumanchipalli, J. Chartier, and E. F. Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

T. Arias-Vergara. *Analysis of pathological speech signals*. PhD thesis, Technische Fakultat der Friedrich-Alexander-Universitat, 2022.

P. Armeni, L. Borsoi, G. Donin, F. Costa, and L. Ferini-Strambi. Pnd33 the clinical and economic burden of obstructive sleep apnea in adults: A cost-of-illness analysis. *Value in Health*, 22:S742, 2019.

E. G. Arnold, G. Lotha, R. Pallardy, and K. Rogers. Speech. *Encyclopædia Britannica*, 2019. URL https://www.britannica.com/topic/speech-language.

J. Arnold, M. Sunilkumar, V. Krishna, S. Yoganand, M. S. Kumar, and D. Shanmugapriyan. Obstructive sleep apnea. *Journal of pharmacy & bioallied sciences*, 9(Suppl 1):S26, 2017.

M. Asiaee, A. Vahedian-Azimi, S. S. Atashi, A. Keramatfar, and M. Nourbakhsh. Voice quality evaluation in patients with COVID-19: An acoustic analysis. *Journal of Voice*, 2020.

A. P. Association et al. Diagnostic and statistical manual of mental disorders (dsm-5®): American psychiatric pub; 2013. *J. Physiother. Res., Salvador*, 9(2):155–158, 2019.

L. Bäckman, S. Jones, A.-K. Berger, E. J. Laukka, and B. Small. Multiple cognitive deficits during the transition to Alzheimer's disease. *Journal of internal medicine*, 256(3):195–204, 2004.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare, A. Mahale, S. Rane, N. Agarwal, and R. Panicker. Cough against COVID: Evidence of COVID-19 signature in cough sounds. *preprint arXiv:2009.08790*, 2020.

M. H. Bahari, M. McLaren, D. A. van Leeuwen, et al. Speaker age estimation using i-vectors. *Engineering Applications of Artificial Intelligence*, 34:99–108, 2014.

A. Balaei, K. Sutherland, and P. Cistulli et al. Automatic detection of obstructive sleep apnea using facial images. In *ISBI*. IEEE, 2017.

A. T. Balaei, K. Sutherland, P. Cistulli, and P. de Chazal. Prediction of obstructive sleep apnea using facial landmarks. *Physiological measurement*, 39(9):094004, 2018.

J.-U. Bang, S.-H. Han, and B.-O. Kang. Alzheimer's Disease recognition from spontaneous speech using large language models. *ETRI Journal*, 46(1):96–105, 2024. doi: https://doi.org/10.4218/etrij.2023-0356.

K. Barnett, S. W. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, 380(9836):37–43, 2012.

R. Battle and T. Gollapudi. The unreasonable effectiveness of eccentric automatic prompts. *arXiv*, abs/2402.10949, 2024.

A. T. Beck. *Depression: clinical, experimental, and theoretical aspects*. PhD thesis, University of Pennsylvania, 1967.

J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle. The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994. ISSN 0003-9942. doi: 10.1001/archneur.1994.00540180063015.

G. Bedi, F. Carrillo, G. A. Cecchi, D. F. Slezak, M. Sigman, N. B. Mota, S. Ribeiro, D. C. Javitt, M. Copelli, and C. M. Corcoran. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1(1):1–7, 2015.

J. Behrendt. Demvis: Modular system for speech-based dementia screening. Master's thesis, University of Bremen, 2023.

A. M. Benavides, R. F. Pozo, D. T. Toledano, J. L. B. Murillo, E. L. Gonzalo, and L. H. Gómez. Analysis of voice features related to obstructive sleep apnoea and their application in diagnosis support. *Computer Speech & Language*, 28(2):434–452, 2014.

A. V. Benjafield, N. T. Ayas, P. R. Eastwood, R. Heinzer, M. S. Ip, M. J. Morrell, C. M. Nunez, S. R. Patel, T. Penzel, J.-L. Pépin, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *The Lancet Respiratory Medicine*, 7(8):687–698, 2019.

V. Berisha, S. Wang, A. LaCross, and J. Liss. Tracking discourse complexity preceding Alzheimer's disease diagnosis: A case study comparing the press conferences of presidents ronald reagan and george herbert walker bush. *Journal of Alzheimer's Disease*, 45(3):959–963, 2015.

V. Berisha, C. Krantsevich, G. Stegmann, S. Hahn, and J. Liss. Are reported accuracies in the clinical speech machine learning literature overoptimistic? In *Interspeech*, pages 2453–2457, 2022.

P. Boersma. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345, 2001.

P. Boersma and D. Weenink. Praat: doing phonetics by computer [computer program]. Retrieved May 2024 from http://www.praat.org/, 2024.

V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa. Connected speech in neurodegenerative language disorders: a review. *Frontiers in psychology*, 8:208495, 2017.

C. Botelho. Speech as a biomarker for sleep disorders and sleep deprivation. Master's thesis, Instituto Superior Técnico, University of Lisbon, 2018.

C. Botelho, I. Trancoso, A. Abad, and T. Paiva. Speech as a biomarker for obstructive sleep apnea detection. In *ICASSP*, pages 5851–5855. IEEE, 2019.

C. Botelho, L. Diener, D. Küster, K. Scheck, S. Amiriparian, B. W. Schuller, T. Schultz, A. Abad, and I. Trancoso. Toward silent paralinguistics: Speech-to-EMG – retrieving articulatory muscle activity from speech. In *Interspeech*, 2020a.

C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso. Pathological speech detection using x-vector embeddings. *arXiv preprint arXiv:2003.00864*, 2020b.

C. Botelho, A. Abad, T. Schultz, and I. Trancoso. Visual speech for obstructive sleep apnea detection. In *Interspeech*, 2021.

C. Botelho, T. Schultz, A. Abad, and I. Trancoso. Challenges of using longitudinal and cross-domain corpora on studies of pathological speech. In *Interspeech*, 2022.

C. Botelho, A. Abad, T. Schultz, and I. Trancoso. Towards reference speech characterization for health applications. In *Interspeech*, 2023.

C. Botelho, A. Abad, T. Schultz, and I. Trancoso. Speech as a biomarker for disease detection. *IEEE Access*, 12:184487–184508, 2024. doi: 10.1109/ACCESS.2024.3506433.

G. Botha, G. Theron, R. Warren, M. Klopper, K. Dheda, P. Van Helden, and T. Niesler. Detection of tuberculosis by automatic cough sound analysis. *Physiological measurement*, 39(4):045005, 2018.

L. K. Bowen, G. L. Hands, S. Pradhan, and C. E. Stepp. Effects of Parkinson's disease on fundamental frequency variability in running speech. *Journal of medical speech-language pathology*, 21(3):235, 2013.

C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound

data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 3474–3484, Virtual Event, CA, USA, 2020. ISBN 9781450379984. doi: 10.1145/3394486.3412865.

P. Brown, D. Corcos, and J. Rothwell. Does parkinsonian action tremor contribute to muscle weakness in Parkinson's disease? *Brain: a journal of neurology*, 120(3):401–408, 1997.

J. E. Brush Jr, J. Sherbino, and G. R. Norman. How expert clinicians intuitively recognize a medical diagnosis. *The American journal of medicine*, 130(6):629–634, 2017.

G. A. Bryant and M. G. Haselton. Vocal cues of ovulation in human females. *Biology Letters*, 5(1):12–15, 2009.

R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91, 2000.

B. Calabrese, F. Pucci, and M. Sturniolo et al. Automatic detection of obstructive sleep apnea syndrome based on snore signals. In *MAVEBA*, pages 185–188, 2009.

Cambridge University Press. Discourse markers: so, right, okay. https://dictionary.cambridge.org/grammar/british-grammar/discourse-markers-so-right-okay. [Accessed: 2024-07-08].

J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, et al. The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer, 2005.

R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

P.-F. Center. How artificial intelligence is revolutionizing drug discovery, 2023. URL https://blog.petrieflom.law.harvard.edu/2023/03/20/how-artificial-intelligence-is-revolutionizing-drug-discovery/. Accessed: 2024-06-20.

Center for Countering Digital Hate. Ai and eating disorders: How AI is encouraging and exacerbating eating disorders among young users, 2023. URL https://counterhate.com/research/ai-tools-and-eating-disorders/. Accessed: 2024-06-21.

G. Chaudhari, X. Jiang, A. Fakhry, A. Han, J. Xiao, S. Shen, and A. Khanzada. Virufy: Global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough. *preprint arXiv:2011.13320*, 2021.

S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016.

E. A. Coker, A. Stewart, B. Ozer, A. Minchom, L. Pickard, R. Ruddle, S. Carreira, S. Popat, M. O'Brien, F. Raynaud, et al. Individualized prediction of drug response and rational combination therapy in NSCLC using artificial intelligence–enabled studies of acute phosphoproteomic changes. *Molecular cancer therapeutics*, 21(6):1020–1029, 2022.

J. Correia. *In-the-wild detection of speech affecting diseases*. PhD thesis, Carnegie Mellon University - University of Lisbon, 2021.

J. Correia, B. Raj, and I. Trancoso. Querying depression vlogs. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 987–993. IEEE, 2018a.

J. Correia, B. Raj, I. Trancoso, and F. Teixeira. Mining multimodal repositories for speech affecting diseases. In *Interspeech*, 2018b.

J. Correia, F. Teixeira, C. Botelho, I. Trancoso, and B. Raj. The in-the-wild speech medical corpus. In *ICASSP*. IEEE, 2021.

N. Cummins and B. W. Schuller. Five crucial challenges in digital health, 2020.

N. Cummins, J. Epps, M. Breakspear, and R. Goecke. An investigation of depressed speech detection: Features and normalization. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015a.

N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski. Analysis of acoustic space variability in speech affected by depression. *Speech Communication*, 75:27–49, 2015b.

N. Cummins, A. Baird, and B. W. Schuller. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54, 2018.

N. Cummins, J. Dineley, P. Conde, F. Matcham, S. Siddi, F. Lamers, E. Carr, G. Lavelle, D. Leightley, K. White, et al. Multilingual markers of depression in remotely collected speech samples. *preprint*, 2022.

E. Dallé and M. V. Mabandla. Early life stress, depression and Parkinson's disease: a new approach. *Molecular brain*, 11(1):1–13, 2018.

P. de Chazal, P. A. Cistulli, and M. T. Naughton. The future of sleep-disordered breathing: A public health crisis. *Respirology*, 2020.

S. de la Fuente García. *Investigating speech technology for monitoring disease progression in the context of neurodegenerative disease*. PhD thesis, University of Edinburgh, 2021.

L. M. De Lau and M. M. Breteler. Epidemiology of Parkinson's disease. *The Lancet Neurology*, 5(6): 525–535, 2006.

N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak. Language recognition via i-vectors and dimensionality reduction. In *Twelfth annual conference of the international speech communication association*, 2011.

B. Denby and M. Stone. Speech synthesis from real time ultrasound images of the tongue. In *ICASSP*, volume 1, pages I–685. IEEE, 2004.

J. Deng, N. Cummins, M. Schmitt, K. Qian, F. Ringeval, and B. Schuller. Speech-based diagnosis of autism spectrum condition by generative adversarial network representations. In *Proceedings of the 2017 international conference on digital health*, pages 53–57, 2017.

D. Dernoncourt, B. Hanczar, and J.-D. Zucker. Analysis of feature selection stability on high dimension and small sample data. *Computational statistics & data analysis*, 71:681–693, 2014.

S. Deshmukh, M. Al Ismail, and R. Singh. Interpreting glottal flow dynamics for detecting COVID-19 from voice. In *ICASSP*, pages 1055–1059. IEEE, 2021.

G. Deshpande and B. Schuller. An overview on audio, signal, speech, & language processing for COVID-19. *arXiv preprint arXiv:2005.08579*, 2020.

B. Desplanques, J. Thienpondt, and K. Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

V. Despotovic, M. Ismael, M. Cornil, R. M. Call, and G. Fagherazzi. Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results. *Computers in Biology and Medicine*, 138:104944, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.104944.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

L. Diener. *The Impact of Audible Feedback on EMG-to-Speech Conversion*. PhD thesis, University of Bremen, 2021. URL https://www.csl.uni-bremen.de/cms/images/documents/publications/Diener2021Diss.pdf.

L. Diener, M. Janke, and T. Schultz. Direct conversion from facial myoelectric signals to speech using deep neural networks. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2015. doi: 10.1109/IJCNN.2015.7280404.

L. Diener, G. Felsch, M. Angrick, and T. Schultz. Session-independent array-based EMG-to-speech conversion using convolutional neural networks. In *Speech Communication; 13th ITG-Symposium*, pages 1–5. VDE, 2018.

L. Diener, S. Amiriparian, C. Botelho, K. Scheck, D. Küster, I. Trancoso, B. W. Schuller, and T. Schultz. Towards silent paralinguistics: Deriving speaking mode and speaker ID from electromyographic signals. In *Interspeech*, 2020.

J. Dineley, E. Carr, F. Matcham, J. Downs, R. Dobson, T. F. Quatieri, and N. Cummins. Towards robust paralinguistic assessment for real-world mobile health (mHealth) monitoring: an initial study of reverberation effects on speech. *Interspeech*, 2023.

V. Dobrovolskii. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.605.

J. Duffy. Speech motor disorders: substrates, differential diagnosis, and management. *St. Louis: Mosby*, 1995.

O. Elisha, A. Tarasiuk, and Y. Zigel. Automatic detection of obstructive sleep apnea using speech signal analysis. In *Afeka-AVIOS Speech Processing Conference 2012*, 2012.

ELRA Catalogue ID ELRA-S0390. Parallel EM-acoustic English GlobalPhone, ISLRN 910-309-096-5, 2014. URL http://www.islrn.org/resources/910-309-096-523-6/.

F. Espinoza-Cuadros, R. Fernández-Pozo, D. T. Toledano, J. D. Alcázar-Ramírez, E. López-Gonzalo, and L. A. Hernández-Gómez. Speech signal and facial image processing for obstructive sleep apnea assessment. *Computational and mathematical methods in medicine*, 2015, 2015.

F. Espinoza-Cuadros, R. Pozo, D. Toledano, J. Alcázar-Ramírez, E. Gonzalo, and L. Hernandez-Gomez et al. Reviewing the connection between speech and obstructive sleep apnea. *Biomedical engineering online*, 15(1):20, 2016.

European Commission. Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206. Accessed: 2024-06-20.

European Parliament and Council. On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Regulation 2016/679*, April 2016.

F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM, 2010.

F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 835–838, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2404-5. doi: 10.1145/2502081.2502224. URL http://doi.acm.org/10.1145/2502081.2502224.

F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 4 2016. ISSN 1949-3045. doi: 10.1109/TAFFC.2015.2457417. Open access.

H. Fang, S. Tu, J. Sheng, and A. Shao. Depression in sleep disturbance: a review on a bidirectional relationship, mechanisms and treatment. *Journal of cellular and molecular medicine*, 23(4):2324–2332, 2019.

D. R. Feinberg. Parselmouth praat scripts in python, Jan 2022. URL osf.io/6dwr3.

L. Ferrer and P. Riera. Confidence intervals for evaluation in machine learning. URL https://github.com/luferrer/ConfidenceIntervals.

M. Ferro. Parkinson detection through visual speech. Master's thesis, Instituto Superior Técnico, University of Lisbon, 2023.

L. Ferrucci, M. Gonzalez-Freire, E. Fabbri, E. Simonsick, T. Tanaka, Z. Moore, S. Salimi, F. Sierra, and R. de Cabo. Measuring biological aging in humans: A quest. *Aging Cell*, 19(2):e13080, 2020.

E. Fischer and A. M. Goberman. Voice onset time in Parkinson disease. *Journal of Communication Disorders*, 43(1):21–34, 2010.

A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *Journal of psychiatric research*, 27(3):309–319, 1993.

K. E. Forbes, A. Venneri, and M. F. Shanks. Distinct patterns of spontaneous speech deterioration: an early predictor of Alzheimer's disease. *Brain and Cognition*, 48(2-3):356–361, 2002.

A. Fox, P. Monoson, and D. Morgan. Speech dysfunction of obstructive sleep apnea: A discriminant analysis of its descriptors. *Chest*, 96(3):589–595, 1989.

C. Frankenberg, J. Weiner, T. Schultz, M. Knebel, C. Degen, H.-W. Wahl, and J. Schröder. Perplexity – a new predictor of cognitive changes in spoken language? – results of the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE). *Linguistics Vanguard*, 5, 06 2019. ISSN 2199174X. doi: 10.1515/lingvan-2018-0026. URL https://www.degruyter.com/view/j/lingvan.2019.5.issue-s2/lingvan-2018-0026/lingvan-2018-0026.xml. s2.

M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller. audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research*, 18(1):6340–6344, 2017.

A. Gabbay, A. Shamir, and S. Peleg. Visual speech enhancement. *arXiv preprint arXiv:1711.08789*, 2017.

J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue. TIMIT acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*, 1992.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

A. Goberman, C. Coelho, and M. Robb. Phonatory characteristics of parkinsonian speech before and after morning medication: the ON and OFF states. *Journal of communication disorders*, 35(3):217–239, 2002.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society, 1992.

C. G. Goetz, W. Poewe, O. Rascol, C. Sampaio, G. T. Stebbins, C. Counsell, N. Giladi, R. G. Holloway, C. G. Moore, G. K. Wenning, et al. Movement disorder society task force report on the hoehn and

yahr staging scale: status and recommendations the movement disorder society task force on rating scales for Parkinson's disease. *Movement disorders*, 19(9):1020–1028, 2004.

E. Goldshtein, A. Tarasiuk, and Y. Zigel. Automatic detection of obstructive sleep apnea using speech signals. *IEEE Transactions on biomedical engineering*, 58(5):1373–1382, 2011.

P. Gómez-Vilda, A. R. M. Londral, V. Rodellar-Biarge, J. M. Ferrández-Vicente, and M. de Carvalho. Monitoring amyotrophic lateral sclerosis by biomechanical modeling of speech production. *Neurocomputing*, 151:130–138, 2015.

H. Goodglass, E. Kaplan, and S. Weintraub. *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

A. Goyal and Y. Bengio. Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091*, 2020.

J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Citeseer, 2014.

B. S. Greenwald. *Depression in Alzheimer's disease and related dementias*. American Psychiatric Press, Washington, 1995.

J. Grosman. Fine-tuned XLSR-53 large model for speech recognition in English. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english, 2021.

G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *2009 IEEE conference on computer vision and pattern recognition*, pages 112–119. IEEE, 2009.

D. C. Halahakoon, G. Lewis, and J. P. Roiser. Cognitive impairment and depression—cause, consequence, or coincidence? *JAMA psychiatry*, 76(3):239–240, 2019.

H. J. Han, S. BN, L. Qiu, and S. Abdullah. Automatic classification of dementia using text and speech data. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, pages 399–407. Springer, 2022.

J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo. Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data. In *ICASSP*, pages 8328–8332, 2021. doi: 10.1109/ICASSP39728.2021.9414576.

B. Harel, M. Cannizzaro, and P. J. Snyder. Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study. *Brain and cognition*, 56(1):24–29, 2004.

K. M. Harrel and R. W. Dudek. *Anatomy*. Lippincott Illustrated Reviews, 2019.

R. Haulcy and J. Glass. CLAC: A speech corpus of healthy English speakers. In *Interspeech*, pages 2966–2970, 2021. doi: 10.21437/Interspeech.2021-1810.

Y. Hauptman, R. Aloni-Lavi, I. Lapidot, T. Gurevich, Y. Manor, S. Naor, N. Diamant, and I. Opher. Identifying distinctive acoustic and spectral features in Parkinson's disease. In *Interspeech*, pages 2498–2502, 2019.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

T. Head, M. Kumar, H. Nahrstaedt, G. Louppe, and I. Shcherbatyi. Scikit-optimize/scikit-optimize, Oct 2021. URL http://doi.org/10.5281/zenodo.1207017.

K. Hechmi, T. N. Trong, V. Hautamaki, and T. Kinnunen. Voxceleb enrichment for age and gender recognition, 2021. URL https://arxiv.org/abs/2109.13510.

P. Hecker, N. Steckhan, F. Eyben, B. W. Schuller, and B. Arnrich. Voice analysis for neurological disorder recognition–a systematic review and perspective on emerging trends. *Frontiers in Digital Health*, 4, 2022.

C. Herff, G. Johnson, L. Diener, J. Shih, D. Krusienski, and T. Schultz. Towards direct speech synthesis from ECoG: A pilot study. In *EMBC*, pages 1540–1543. IEEE, 2016.

S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. CNN architectures for large-scale audio classification. In *ICASSP*, pages 131–135, 2017.

D. B. Hier, K. Hagenlocker, and A. G. Shindler. Language disintegration in dementia: Effects of etiology and severity. *Brain and language*, 25(1):117–133, 1985.

I. Hoffmann, D. Nemeth, C. D. Dye, M. Pákáski, T. Irinyi, and J. Kálmán. Temporal parameters of spontaneous speech in Alzheimer's disease. *International journal of speech-language pathology*, 12 (1):29–34, 2010.

S. D. Holcomb, W. K. Porter, S. V. Ault, G. Mao, and J. Wang. Overview on DeepMind and its AlphaGo Zero AI . In *Proceedings of the 2018 international conference on big data and education*, pages 67–71, 2018.

R. B. Hoodin and H. R. Gilbert. Nasal airflows in parkinsonian speakers. *Journal of Communication Disorders*, 22(3):169–180, 1989.

G. L. Horowitz, S. Altaie, J. Boyd, F. Ceriotti, U. Garg, P. Horn, A. Pesce, E. Harrison, and J. Zakowski. EP28-A3C defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline. *San Diego: Clinical and Laboratory Standards Institute*, 2010.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021a.

W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. In *Proc. Interspeech 2021*, pages 721–725, 2021b. doi: 10.21437/Interspeech.2021-236.

S. Hu, X. Xie, Z. Jin, M. Geng, Y. Wang, M. Cui, J. Deng, X. Liu, and H. Meng. Exploring self-supervised pre-trained ASR models for dysarthric and elderly speech recognition. In *ICASSP*, pages 1–5. IEEE, 2023.

G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

X. Huang, A. Acero, H.-W. Hon, and R. Reddy. *Spoken language processing: A guide to theory, algorithm, and system development*, volume 1. Prentice hall PTR Upper Saddle River, 2001.

K. Ichihara, Y. Itoh, and C. W. Lam et al. Sources of variation of commonly measured serum analytes in 6 Asian cities and consideration of common reference intervals. *Clinical chemistry*, 54(2):356–365, 2008.

K. Ichihara, J. C. Boyd, et al. An appraisal of statistical procedures used in derivation of reference intervals. *Clinical chemistry and laboratory medicine*, 48(11):1537–1551, 2010.

A. Illa and P. Ghosh. Representation learning using convolution neural network for acoustic-to-articulatory inversion. In *ICASSP*, pages 5931–5935, 2019.

A. Illa and P. K. Ghosh. Low resource acoustic-to-articulatory inversion using bi-directional long short term memory. In *Interspeech*, pages 3122–3126, 2018.

S. Imai. Cepstral analysis synthesis on the mel frequency scale. In *ICASSP*, volume 8, pages 93–96, 1983.

A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Informatics in Medicine Unlocked*, 20:100378, 2020. ISSN 2352-9148.

D. Iter, J. Yoon, and D. Jurafsky. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146, 2018.

Y. Jadoul, B. Thompson, and B. de Boer. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15, 2018. doi: https://doi.org/10.1016/j.wocn.2018.07.001.

S. I. Jang, M. Lee, J. Han, J. Kim, A. R. Kim, J. S. An, J. O. Park, B. J. Kim, and E. Kim. A study of skin characteristics with long-term sleep restriction in korean women in their 40s. *Skin Research and Technology*, 26(2):193–199, 2020.

M. Janke and L. Diener. EMG-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25(12):2375–2385, 2017. doi: 10.1109/TASLP.2017.2738568.

S. Jannetts, F. Schaeffler, J. Beck, and S. Cowen. Assessing voice health using smartphones: bias and random error of acoustic voice parameters captured by different smartphone types. *International journal of language & communication disorders*, 54(2):292–305, 2019.

W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *CLPsych*, 2014.

L. Jeancolas, D. Petrovska-Delacrétaz, G. Mangone, B.-E. Benkelfat, J.-C. Corvol, M. Vidailhet, S. Lehéricy, and H. Benali. x-vectors: New quantitative biomarkers for early Parkinson's disease detection from speech. *Frontiers in Neuroinformatics*, 15:4, 2021.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7B. *arXiv*, abs/2310.06825, 2023.

A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mixtral of experts. *arXiv*, abs/2401.04088, 2024.

Y. Jiao, V. Berisha, and J. Liss. Interpretable phonological features for clinical applications. In *ICASSP*, pages 5045–5049. IEEE, 2017.

F. J. Jiménez-Jiménez, J. Gamboa, A. Nieto, J. Guerrero, M. Orti-Pareja, J. A. Molina, E. García-Albea, and I. Cobeta. Acoustic voice analysis in untreated patients with Parkinson's disease. *Parkinsonism & Related Disorders*, 3(2):111–116, 1997.

G. R. Jones, R. Haeckel, T. P. Loh, K. Sikaris, T. Streichert, A. Katayev, J. H. Barth, Y. Ozarda, et al. Indirect methods for reference interval determination–review and recommendations. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 57(1):20–29, 2018.

S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel. Towards continuous speech recognition using surface electromyography. In *Ninth International Conference on Spoken Language Processing*, 2006.

S.-C. S. Jou. *Automatic speech recognition on vibrocervigraphic and electromyographic signals*. PhD thesis, Carnegie Mellon University, Language Technologies Institute, 2008.

D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

S. Kapoor and A. Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 2023.

A. Kappas, E. Krumhuber, and D. Küster. Facial behavior. In *In: Hall, Judith A.; Knapp, Mark L. (Ed.), Nonverbal communication (pp. 131-166). Berlin: de Gruyter, 2013*, pages 131–166. de Gruyter. ISBN 978-3-11-023814-3.

Z. N. Karam, E. M. Provost, S. Singh, J. Montgomery, C. Archer, G. Harrington, and M. G. Mcinnis. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In *ICASSP*, pages 4858–4862. IEEE, 2014.

M. R. Kaufman, E. L. Eschliman, and T. S. Karver. Differentiating sex and gender in health research to achieve gender equity. *Bulletin of the World Health Organization*, 101(10):666, 2023.

V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

J. G. Kerns and H. Berenbaum. Cognitive impairments associated with formal thought disorder in people with schizophrenia. *Journal of abnormal psychology*, 111(2):211, 2002.

M. A. Kim, E. J. Kim, B. Y. Kang, and H. K. Lee. The effects of sleep deprivation on the biophysical properties of facial skin. *Journal of Cosmetics, Dermatological Sciences and Applications*, 7(1):34–47, 2017.

N. Kimura, M. Kono, and J. Rekimoto. SottoVoce: an ultrasound imaging-based silent speech interaction using deep neural networks. In *2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

D. King. High quality face recognition with deep metric learning, http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html, Feb 2017. URL http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html.

R. Kliper, S. Portuguese, and D. Weinshall. Prosodic analysis of speech and the underlying mental state. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, pages 52–62. Springer, 2015.

M. Kriboy, A. Tarasiuk, and Y. Zigel. Detection of obstructive sleep apnea in awake subjects by exploiting body posture effects on the speech signal. In *EMBC*. IEEE, 2014a.

M. Kriboy, A. Tarasiuk, and Y. Zigel. A novel method for obstructive sleep apnea severity estimation using speech signals. In *ICASSP*. IEEE, 2014b.

G. Krishna, Y. Han, C. Tran, M. Carnahan, and A. H. Tewfik. State-of-the-art speech recognition using EEG and towards decoding of speech spectrum from EEG. *arXiv preprint arXiv:1908.05743*, 2019.

G. Krishna, C. Tran, M. Carnahan, Y. Han, and A. H. Tewfik. Generating EEG features from acoustic features. *arXiv preprint arXiv:2003.00007*, 2020a.

G. Krishna, C. Tran, Y. Han, and M. Carnahan. Speech synthesis using EEG. *arXiv preprint arXiv:2002.12756*, 2020b.

K. R. R. Krishnan, M. Delong, H. Kraemer, R. Carney, D. Spiegel, C. Gordon, W. McDonald, M. A. Dew, G. Alexopoulos, K. Buckwalter, et al. Comorbidity of depression with other medical diseases in the elderly. *Biological psychiatry*, 52(6):559–588, 2002.

K. Kroenke and R. L. Spitzer. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric annals*, 32(9):509–515, 2002.

D. Küster. Hidden tears and scrambled joy: On the adaptive costs of unguarded nonverbal social signals. In *Social Intelligence and Nonverbal Communication*, pages 283–304. Springer, 2020.

D. Küster and T. Schultz. Künstliche intelligenz und ethik im gesundheitswesen–spagat oder symbiose? *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz*, 66(2):176–183, 2023.

D. Kwasny and D. Hemmerling. Gender and age estimation methods based on speech using deep neural networks. *Sensors*, 21(14):4785, 2021.

A. Lahti, P. H. Petersen, J. C. Boyd, C. G. Fraser, and N. Jørgensen. Objective criteria for partitioning gaussian-distributed reference values into subgroups. *Clinical chemistry*, 48(2):338–352, 2002.

K. T. Laird, B. Krause, C. Funes, and H. Lavretsky. Psychobiological factors of resilience and depression in late life. *Translational Psychiatry*, 9(1):1–18, 2019.

J. Laver. *Principles of phonetics*. Cambridge university press, 1994.

J. R. Lechien, C. M. Chiesa-Estomba, P. Cabaraux, Q. Mat, K. Huet, B. Harmegnies, M. Horoi, S. D. Le Bon, A. Rodriguez, D. Dequanter, et al. Features of mild-to-moderate COVID-19 patients with dysphonia. *Journal of Voice*, 2020.

R. W. Lee, A. S. Chan, R. R. Grunstein, and P. A. Cistulli. Craniofacial phenotyping in obstructive sleep apnea—a novel quantitative photographic approach. *Sleep*, 32(1):37–45, 2009a.

R. W. Lee, P. Petocz, T. Prvan, A. S. Chan, R. R. Grunstein, and P. A. Cistulli. Prediction of obstructive sleep apnea with craniofacial photographic analysis. *Sleep*, 32(1):46–52, 2009b.

S. Lee, J. Park, and D. Um. Speech characteristics as indicators of personality traits. *Applied Sciences*, 11(18):8776, 2021.

L. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai. A deep recurrent approach for acoustic-to-articulatory inversion. In *ICASSP*, pages 4450–4454. IEEE, 2015.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

Y.-L. Liu, R. Feng, J.-H. Yuan, and Z.-H. Ling. Clever Hans effect found in automatic detection of Alzheimer's disease through speech. *arXiv preprint arXiv:2406.07410*, 2024.

K. Lopez-de Ipina, U. Martinez-de Lizarduy, P. M. Calvo, J. Mekyska, B. Beitia, N. Barroso, A. Estanga, M. Tainta, and M. Ecay-Torres. Advances on automatic speech analysis for early detection of Alzheimer's disease: a non-linear multi-task approach. *Current Alzheimer Research*, 15(2):139–148, 2018.

P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo. i-vectors for continuous emotion recognition. *Training*, 45:50, 2014.

P. Lopez-Otero, L. Docio-Fernandez, A. Abad, and C. Garcia-Mateo. Depression detection using automatic transcriptions of de-identified speech. *Interspeech*, pages 3157–3161, 2017.

S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge. *Interspeech*, 2020.

S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney. Detecting cognitive decline using speech only: The ADReSSo Challenge. *arXiv preprint arXiv:2104.09356*, 2021.

M. M. Lyons, N. Y. Bhatt, A. I. Pack, and U. J. Magalang. Global burden of sleep-disordered breathing and its implications. *Respirology*, 25(7):690–702, 2020.

A. Ma, K. K. Lau, and D. Thyagarajan. Voice changes in Parkinson's disease: What are they telling us? *Journal of Clinical Neuroscience*, 72:1–7, 2020.

P. Ma, B. Martinez, S. Petridis, and M. Pantic. Towards practical lipreading with distilled and efficient models. In *ICASSP*. IEEE, 2021.

Z. Ma, Y. Mei, and Z. Su. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1105. American Medical Informatics Association, 2023.

L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008.

C. Mackenzie, M. Brady, J. Norrie, and N. Poedjianto. Picture description in neurologically normal adults: Concepts and topic coherence. *Aphasiology*, 21(3-4):340–354, 2007.

P. C. Mahalanobis. On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 80:S1–S7, 2018.

A. Malhotra and D. P. White. Obstructive sleep apnoea. *The lancet*, 360(9328):237–245, 2002.

V. Marian, J. Bartolotti, S. Chabal, and A. Shook. Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. 2012.

P. Martin, M. Grünendahl, and M. Schmitt. Persönlichkeit, kognitive leistungsfähigkeit und gesundheit in ost und west: Ergebnisse der interdisziplinären längsschnittstudie des erwachsenenalters (ilse). *Zeitschrift für Gerontologie und Geriatrie*, 33(2):111–123, 2000.

F. Martínez-Sánchez, J. Meilán, J. Carro, C. G. Íñiguez, L. Millian-Morell, I. P. Valverde, T. López-Alburquerque, and D. Lopez. Speech rate in Parkinson's disease: a controlled study. *Neurologia (English Edition)*, 31(7):466–472, 2016.

Y. Maryn, F. Ysenbaert, A. Zarowski, and R. Vanspauwen. Mobile communication devices, ambient noise, and acoustic voice measures. *Journal of Voice*, 31(2):248–e11, 2017.

S. McGregor. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. *arXiv preprint arXiv:2011.08512*, 2020.

G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia*, 7(3):263–269, 2011.

P. E. McKnight and J. Najab. Mann-Whitney U Test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

H. Meinedo and J. Neto. A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ann models. In *Interspeech*, 2005.

D. Michelsanti, O. Slizovskaia, G. Haro, E. Gómez, Z.-H. Tan, and J. Jensen. Vocoder-based speech synthesis from silent videos. *arXiv preprint arXiv:2004.02541*, 2020.

I. Midi, M. Dogan, M. Koseoglu, G. Can, M. Sehitoglu, and D. Gunal. Voice abnormalities and their relation with motor dysfunction in Parkinson's disease. *Acta Neurologica Scandinavica*, 117(1):26–34, 2008.

M. Milling, F. B. Pokorny, K. D. Bartl-Pokorny, and B. W. Schuller. Is speech the new blood? recent progress in AI-based disease detection from audio in a nutshell. *Frontiers in digital health*, 4:886615, 2022.

P. K. Monoson and A. W. Fox. Preliminary observation of speech disorder in obstructive and mixed sleep apnea. *Chest*, 92(4):670–675, 1987.

L. Moro-Velazquez, J. Villalba, and N. Dehak. Using x-vectors to automatically detect Parkinson's disease from speech. In *ICASSP*, pages 1155–1159. IEEE, 2020.

A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*, 2019.

H. Nam, S.-H. Kim, and Y.-H. Park. Filteraugment: An acoustic environmental data augmentation method. In *ICASSP*, pages 4308–4312. IEEE, 2022.

M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 43–50, 2016.

H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*, pages 343–347. IEEE, 2014.

L. E. Nicholas and R. H. Brookshire. Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech, Language, and Hearing Research*, 38(1):145–156, 1995.

G. Noffs, T. Perera, S. C. Kolbe, C. J. Shanahan, F. M. Boonstra, A. Evans, H. Butzkueven, A. van der Walt, and A. P. Vogel. What speech can tell us: A systematic review of dysarthria characteristics in multiple sclerosis. *Autoimmunity reviews*, 17(12):1202–1209, 2018.

G. Noffs, F. Boonstra, T. Perera, H. Butzkueven, S. Kolbe, et al. Speech metrics, general disability, brain imaging and quality of life in multiple sclerosis. *European Journal of Neurology*, 28(1):259–268, 2021.

H. Nosrati, N. Sadr, and P. de Chazal. Apnoea-hypopnoea index estimation using craniofacial photographic measurements. In *CinC*. IEEE, 2016.

M. D. S. T. F. on Rating Scales for Parkinson's Disease. The unified Parkinson's disease rating scale (UPDRS): status and recommendations. *Movement Disorders*, 18(7):738–750, 2003.

J. C. Ong and M. R. Crawford. Insomnia and obstructive sleep apnea. *Sleep medicine clinics*, 8(3): 389–398, 2013.

C. Opdebeeck, C. Quinn, S. M. Nelis, and L. Clare. Does cognitive reserve moderate the association between mood and cognition? a systematic review. *Reviews in Clinical Gerontology*, 25(3):181–193, 2015.

OpenAI. Introducing GPT-4o, 2024. URL https://openai.com/index/hello-gpt-4o/. Accessed: 2024-06-20.

G. Oppenheim. The earliest signs of Alzheimer's disease. *Journal of geriatric psychiatry and neurology*, 7(2):116–120, 1994.

L. Orlandic, T. Teijeiro, and D. Atienza. The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms. *preprint arXiv:2009.11644*, 2020.

J. Orozco-Arroyave, F. Hönig, J. Arias-Londoño, J. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Nöth. Automatic detection of Parkinson's disease in running speech spoken in three different languages. *The Journal of the Acoustical Society of America*, 139(1):481–500, 2016.

J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth. New spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *LREC*, pages 342–347, 2014.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

Y. Ozarda. Reference intervals: current status, recent developments and future considerations. *Biochemia medica: Biochemia medica*, 26(1):5–16, 2016.

Y. Ozarda, K. Sikaris, T. Streichert, J. Macri, I. C. on Reference intervals, and D. L. (C-RIDL). Distinguishing reference intervals and clinical decision limits–A review by the IFCC Committee on reference intervals and decision limits. *Critical reviews in clinical laboratory sciences*, 55(6):420–431, 2018.

A. S. Ozbolt, L. Moro-Velazquez, I. Lina, A. A. Butala, and N. Dehak. Things to consider when automatically detecting Parkinson's disease using the phonation of sustained vowels: analysis of methodological issues. *Applied Sciences*, 12(3):991, 2022.

T. Paiva, M. Andersen, and S. Tufik. Sono e a medicina do sono. 1ª edição, 2014.

A. Pal and M. Sankarasubbu. Pay attention to the cough: Early diagnosis of COVID-19 using interpretable symptoms embeddings with cough sound signal processing. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 620–628, 2021. ISBN 9781450381048.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an ASR corpus based on public domain audio books. In *ICASSP*, pages 5206–5210, 2015.

D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015.

S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio. Learning problem-agnostic speech representations from multiple self-supervised tasks. In *Interspeech*, 2019.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Penguin Books, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. M. Perero-Codosero, F. Espinoza-Cuadros, J. Antón-Martín, M. A. Barbero-Álvarez, and L. A. Hernández-Gómez. Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training. *IEEE Journal of Selected Topics in Signal Processing*, 14(2): 240–250, 2019.

G. Pinkas, Y. Karny, A. Malachi, G. Barkai, G. Bachar, and V. Aharonson. SARS-CoV-2 Detection From Voice. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:268–274, 2020.

A. Pompili, A. Abad, P. Romano, I. P. Martins, R. Cardoso, H. Santos, J. Carvalho, I. Guimarães, and J. J. Ferreira. Automatic detection of Parkinson's disease: An experimental analysis of common speech production tasks used for diagnosis. In *International Conference on Text, Speech, and Dialogue*, pages 411–419. Springer, 2017.

A. Pompili, A. Abad, D. M. de Matos, and I. P. Martins. Pragmatic aspects of discourse production for the automatic identification of Alzheimer's disease. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):261–271, 2020a. doi: 10.1109/JSTSP.2020.2967879.

A. Pompili, T. Rolland, and A. Abad. The inesc-id multi-modal system for the adress 2020 challenge. In *Interspeech*, 2020b.

A. Pompili, R. Solera-Urena, A. Abad, R. Cardoso, I. Guimaraes, M. Fabbri, I. P. Martins, and J. Ferreira. Assessment of Parkinson's disease medication state through automatic speech analysis. *arXiv preprint arXiv:2005.14647*, 2020c.

A. M. Pompili. *Speech and language technologies applied to diagnostics and therapy of brain diseases*. PhD thesis, Instituto SUperior Técnico - University of Lisbon, 2019.

D. Porras, A. Sepúlveda-Sepúlveda, and T. G. Csapó. DNN-based acoustic-to-articulatory inversion using ultrasound tongue imaging. In *IJCNN*, pages 1–8. IEEE, 2019.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi Speech Recognition Toolkit. In *ASRU*, Dec. 2011.

D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Interspeech*, pages 3743–3747, Hyderabad, India, 2018.

R. F. Pozo, J. L. B. Murillo, L. H. Gómez, E. L. Gonzalo, J. A. Ramírez, and D. T. Toledano. Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques. *EURASIP Journal on Advances in Signal Processing*, 2009(1):982531, 2009.

S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-4501.

R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas. A cough-based algorithm for automatic diagnosis of pertussis. *PLOS ONE*, 11(9):1–20, Sept. 2016.

N. Punjabi. The epidemiology of adult obstructive sleep apnea. *Proceedings of the American Thoracic Society*, 5(2):136–143, 2008.

T. Pyszczynski and J. Greenberg. Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression. *Psychological bulletin*, 102(1):122, 1987.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. of the 40th International Conference on Machine Learning*, pages 28492–28518, 2023.

L. O. Ramig, C. Fox, and S. Sapir. Speech treatment for Parkinson's disease. *Expert Review of Neurotherapeutics*, 8(2):297–309, 2008.

M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio. Multi-task self-supervised learning for robust speech recognition. *preprint ArXiv:2001.09239*, 2020.

M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

V. Ravi, J. Wang, J. Flint, and A. Alwan. Fraug: A frame rate based data augmentation method for depression detection from speech signals. In *ICASSP*, pages 6267–6271. IEEE, 2022.

P. Refaeilzadeh, L. Tang, and H. Liu. Cross-validation. *Encyclopedia of database systems*, 5:532–538, 2009.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

M. Riley, E. Heinen, and J. Ghosh. A text retrieval approach to content-based audio retrieval. In *Int. Symp. on Music Information Retrieval (ISMIR)*, pages 295–300, 2008.

W. G. Rosen, R. C. Mohs, and K. L. Davis. A new rating scale for Alzheimer's disease. *The American journal of psychiatry*, 1984.

S. Rude, E.-M. Gortner, and J. Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.

M. Sagar and R. Scott. System and method for tracking facial muscle and eye motion for computer graphics animation, June 30 2009. US Patent 7,554,549.

C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

C. Sanz, F. Carrillo, A. Slachevsky, G. Forno, M. L. Gorno Tempini, R. Villagra, A. Ibáñez, E. Tagliazucchi, and A. M. García. Automated text-level semantic markers of Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 14(1):e12276, 2022.

C. Sattler, H.-W. Wahl, J. Schröder, A. Kruse, P. Schönknecht, U. Kunzmann, and A. Zenthöfer. Interdisciplinary longitudinal study on adult development and aging (ILSE). *Encyclopedia of geropsychology*, pages 1–10, 2015.

M. Saxon, J. Liss, and V. Berisha. Objective measures of plosive nasalization in hypernasal speech. In *ICASSP*, pages 6520–6524. IEEE, 2019.

K. Scheck and T. Schultz. Ste-gan: Speech-to-electromyography signal conversion using generative adversarial networks. In *Interspeech*, pages 1174–1178, 2023.

H. Schmid. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer, 1999.

H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*, page 154, 2013.

M. Schmitt, F. Ringeval, and B. Schuller. At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. 2016.

B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski. The interspeech 2011 speaker state challenge. In *Interspeech*, 2011.

B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1): 4–39, 2013a.

B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Interspeech*, 2013b.

B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger. The interspeech 2015 computational paralinguistics challenge: nativeness, Parkinson's & eating condition. In *Sixteenth annual conference of the international speech communication association*, 2015.

B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, et al. The Interspeech 2020 computational paralinguistics challenge: elderly emotion, breathing & masks. In *Interspeech*, 2020.

B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, L. J. M. Rothkrantz, J. Zwerts, J. Treep, and C. Kaandorp. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. In *Interspeech*, Brno, Czechia, Sept. 2021.

T. Schultz, M. Wand, T. Hueber, K. D. J., C. Herff, and J. S. Brumberg. Biosignal-based spoken communication: A survey. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25(12): 2257–2271, 2017. doi: 10.1109/TASLP.2017.2752365. URL https://www.csl.uni-bremen.de/cms/images/documents/publications/TASLP-2017-biosignal-based-spoken.pdf.

M. Schünke, E. Schulte, and U. Schumacher. *Prometheus-Lernatlas der Anatomie*. Stuttgart, New York: Thieme Verlag, 2006.

J. W. Schwoebel, J. Schwartz, L. A. Warrenburg, R. Brown, A. Awasthi, A. New, M. Butler, M. Moss, and E. K. Pissadaki. A longitudinal normative dataset and protocol for speech and language biomarker research. *medrxiv*, pages 2021–08, 2021.

A. Shah, H. Dhamyal, Y. Gao, R. Singh, and B. Raj. On the pragmatism of using binary classifiers over data intensive neural network classifiers for detection of COVID-19 from voice. *Interspeech*, 2022.

N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, P. K. Ghosh, S. Ganapathy, et al. Coswara–a database of breathing, cough, and voice sounds for COVID-19 diagnosis. *arXiv preprint arXiv:2005.10548*, 2020.

A. Shivkumar, J. Weston, R. Lenain, and E. Fristed. Blabla: Linguistic feature extraction for clinical analysis in multiple languages. In *Interspeech*, 2020.

J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. d. C. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv. Towards learning a universal non-semantic representation of speech. *Interspeech*, 2021.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, 2015.

R. Singh. *Profiling humans from their voice*, volume 41. Springer, 2019.

K. R. Sitek, D. H. Mathalon, B. J. Roach, J. F. Houde, C. A. Niziolek, and J. M. Ford. Auditory cortex processes variation in our own speech. *PloS one*, 8(12):e82925, 2013.

S. Skodda. Aspects of speech rate and regularity in Parkinson's disease. *Journal of the neurological sciences*, 310(1-2):231–236, 2011.

D. A. Snowdon, S. J. Kemper, J. A. Mortimer, L. H. Greiner, D. R. Wekstein, and W. R. Markesbery. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the nun study. *Jama*, 275(7):528–532, 1996.

D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, pages 999–1003, 2017a.

D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, pages 999–1003, 2017b.

D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. x-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, pages 5329–5333. IEEE, 2018.

D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. x-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, pages 5329–5333, April 2018. doi: 10.1109/ICASSP.2018.8461375.

I. Soares, J. Dias, H. Rocha, M. do Carmo Lopes, and B. Ferreira. Feature selection in small databases: a medical-case study. In *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016: MEDICON 2016, March 31st-April 2nd 2016, Paphos, Cyprus*, pages 814–819. Springer, 2016.

J. Solé-Casals, C. Munteanu, and O. Martín et al. Detection of severe obstructive sleep apnea through voice analysis. *Applied Soft Computing*, 23:346–354, 2014.

R. Solera-Ureña, C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso. Transfer learning-based cough representations for automatic detection of COVID-19. In *Interspeech*, 2021.

W.-J. Song, C. K. Hui, J. H. Hull, S. S. Birring, L. McGarvey, S. B. Mazzone, and K. F. Chung. Confronting COVID-19-associated cough and the post-covid syndrome: role of viral neurotropism, neuroinflammation, and neuroimmune responses. *The Lancet Respiratory Medicine*, 9(5):533–544, 2021.

G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379, 2020.

M. C. Stöppler. Definition of subclinical disease (accessed on july 15, 2022). https://www.rxlist.com/subclinical_disease/definition.htm, 2021.

C. Suess and R. Hausmann. Gross and histopathological pulmonary findings in a COVID-19 associated death during self-isolation. *International journal of legal medicine*, 134(4):1285–1290, 2020.

F. Sullivan. Hidden health crisis costing america billions: Underdiagnosing and undertreating obstructive sleep apnea draining healthcare system. *American Academy of Sleep Medicine*, 2016.

K. Sutherland, R. W. Lee, and P. A. Cistulli. Obesity and craniofacial structure as risk factors for obstructive sleep apnoea: impact of ethnicity. *Respirology*, 17(2):213–222, 2012.

M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova. Automated screening for Alzheimer's dementia through spontaneous speech. In *Interspeech*, 2020.

K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu. Audio-visual speech separation and dereverberation with a two-stage multimodal network. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):542–553, 2020.

Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

J. P. Teixeira and P. O. Fernandes. Jitter, shimmer and HNR classification within gender, tones and vowels in healthy voices. *Procedia technology*, 16:1228–1237, 2014.

K. B. Tølbøll. Linguistic features in depression: a meta-analysis. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, 4(2):39–59, 2019.

L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, and G. Szatlóczki. Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In *Interspeech*. ISCA, 2015.

H. Touvron, L. Martin, K. Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, abs/2307.09288, 2023.

G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *ICASSP*, pages 5200–5204. IEEE, 2016.

M. Tu, V. Berisha, and J. Liss. Interpretable objective assessment of dysarthric speech based on deep neural networks. In *Interspeech*, pages 1849–1853, 2017.

S. Ullah and D.-H. Kim. An optimized EMG encoder to minimize soft speech loss for speech to EMG conversions. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 215–218. IEEE, 2024.

S. S. Upadhya, A. Cheeran, and J. Nirmal. Statistical comparison of jitter and shimmer voice features for healthy and Parkinson affected persons. In *2017 second international conference on electrical, computer and communication technologies (ICECCT)*, pages 1–6. IEEE, 2017.

A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson. Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11):e0224365, 2019.

M. F. Valstar, J. Gratch, B. W. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016 - depression, mood, and emotion recognition workshop and challenge. *CoRR*, abs/1605.01600, 2016. URL http://arxiv.org/abs/1605.01600.

J. Vanek, J. Prasko, S. Genzor, M. Ociskova, K. Kantor, M. Holubova, M. Slepecky, V. Nesnidal, A. Kolek, and M. Sova. Obstructive sleep apnea, depression and cognitive impairment. *Sleep medicine*, 72: 50–58, 2020.

G. Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, 180: 68–77, 2018.

J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth. Convolutional neural network to model articulation impairments in patients with Parkinson's disease. In *Interspeech*, pages 314–318, 2017.

J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, and N. Dehak. State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations. *Computer Speech & Language*, 60:101026, 2020. ISSN 0885-2308.

J. M. Vojtech, C. L. Mitchell, L. Raiff, J. C. Kline, and G. De Luca. Prediction of voice fundamental frequency and intensity from surface electromyographic signals of the face and neck. *Vibration*, 5(4): 692–710, 2022.

R. Voleti, J. M. Liss, and V. Berisha. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE journal of selected topics in signal processing*, 14(2): 282–298, 2019.

K. Vougioukas, P. Ma, S. Petridis, and M. Pantic. Video-driven speech reconstruction using generative adversarial networks. *arXiv preprint arXiv:1906.06301*, 2019.

N. Vyas, S. Saxena, and T. Voice. Learning soft labels via meta learning. *arXiv preprint arXiv:2009.09496*, 2020.

M. Wand, M. Janke, and T. Schultz. The EMG-UKA corpus for electromyographic speech processing. In *Interspeech*, 2014. URL trialdataathttp://www.csl.uni-bremen.de/CorpusData/download.php?crps=EMG.

M. Wand, T. Schultz, and J. Schmidhuber. Domain-adversarial training for session independent EMG-based speech recognition. In *Interspeech*, pages 3167–3171, 2018.

H. H. Wang, J. J. Wang, S. Y. Wong, M. C. Wong, F. J. Li, P. X. Wang, Z. H. Zhou, C. Y. Zhu, S. M. Griffiths, and S. W. Mercer. Epidemiology of multimorbidity in china and implications for the healthcare system: cross-sectional survey among 162,464 community household residents in southern china. *BMC medicine*, 12(1):1–12, 2014.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

J. Weiner and T. Schultz. Automatic screening for transition into dementia using speech. In *ITG*, 2018. URL https://www.csl.uni-bremen.de/cms/images/documents/publications/ITG2018_WeinerEtAl.pdf.

J. Weiner, C. Herff, and T. Schultz. Speech-based detection of Alzheimer's disease in conversational german. In *Interspeech*, pages 1938–1942, 2016.

WHO. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia, September 2021 (accessed on September 28, 2021).

WHO. Coronavirus disease (COVID-19), 2022a. URL https://www.who.int/health-topics/coronavirus#tab=tab_2.

WHO. Depression, 2022b. URL https://www.who.int/news-room/fact-sheets/detail/depression.

WHO. Ageing, (access date: April 16, 2023. URL https://www.who.int/health-topics/ageing#tab=tab_1.

R. S. Wilson, A. W. Capuano, P. A. Boyle, G. M. Hoganson, L. P. Hizel, R. C. Shah, S. Nag, J. A. Schneider, S. E. Arnold, and D. A. Bennett. Clinical-pathologic study of depressive symptoms and cognitive decline in old age. *Neurology*, 83(8):702–709, 2014.

J. Wiseman. py-WebRTCVAD, retrieved in March 2021. URL https://github.com/wiseman/py-webrtcvad.

World Health Organization. Multimorbidity: Technical series on safer primary care, 2016. URL https://apps.who.int/iris/bitstream/handle/10665/252275/9789241511650-eng.pdf.

V. Wurcel, A. Cicchetti, L. Garrison, M. M. Kip, H. Koffijberg, A. Kolbe, M. M. Leeflang, T. Merlin, J. Mestre-Ferrandiz, W. Oortwijn, et al. The value of diagnostic information in personalised healthcare: A comprehensive concept to facilitate bringing this technology into healthcare systems. *Public health genomics*, 22(1-2):8–15, 2019.

S. Yang, F. Wang, L. Yang, F. Xu, M. Luo, X. Chen, X. Feng, and X. Zou. The physical significance of acoustic parameters and its clinical significance of dysarthria in Parkinson's disease. *Scientific Reports*, 10(1):11776, 2020.

Y. Yang, Z. Song, J. Zhuo, M. Cui, J. Li, B. Yang, Y. Du, Z. Ma, X. Liu, Z. Wang, et al. GigaSpeech 2: An evolving, large-scale and multi-domain ASR corpus for low-resource languages with automated crawling, transcription and refinement. *arXiv preprint arXiv:2406.11546*, 2024.

G. Yoon, J. Kramer, A. Zanko, M. Guzijan, S. Lin, A. Foster-Barber, and A. Boxer. Speech and language delay are early manifestations of juvenile-onset huntington disease. *Neurology*, 67(7):1265–1267, 2006.

J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church. Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. In *Interspeech*, 2020.

S. Zargarbashi and B. Babaali. A multi-modal feature embedding approach to diagnose Alzheimer disease from spoken language. *arXiv preprint arXiv:1910.00330*, 2019.

A. Zhang. Speech recognition (version 3.8)[software]. In *Proceedings of ICCC*, 2017.

X. Zhou, K. Jin, Y. Shang, and G. Guo. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*, 11(3):542–552, 2018.

Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen. A review of recent advances in visual speech decoding. *Image and vision computing*, 32(9):590–605, 2014.

F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

M. Zusag, L. Wagner, and T. Bloder. Careful Whisper - leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification. In *Interspeech*, pages 3013–3017, 2023. doi: 10.21437/Interspeech.2023-1653.

A

Appendix: Disease categorization according to the ICD-11

The diagram in Fig. A.1 presents the categories and codes of the speech affecting diseases described in section 2.2, according to the ICD-11[1]. It is clear that the diseases described in Fig. 2.5 are not at the same hierarchical level in Fig. A.1. In fact, with the current available knowledge, one could argue that we should discuss speech affecting diseases in groups of diseases, or higher hierarchical levels. However, we frequently discuss the diseases for which we have labelled data available.

[1] International Classification of Diseases 11th Revision (ICD-11) by the World Health Organization, https://icd.who.int/en

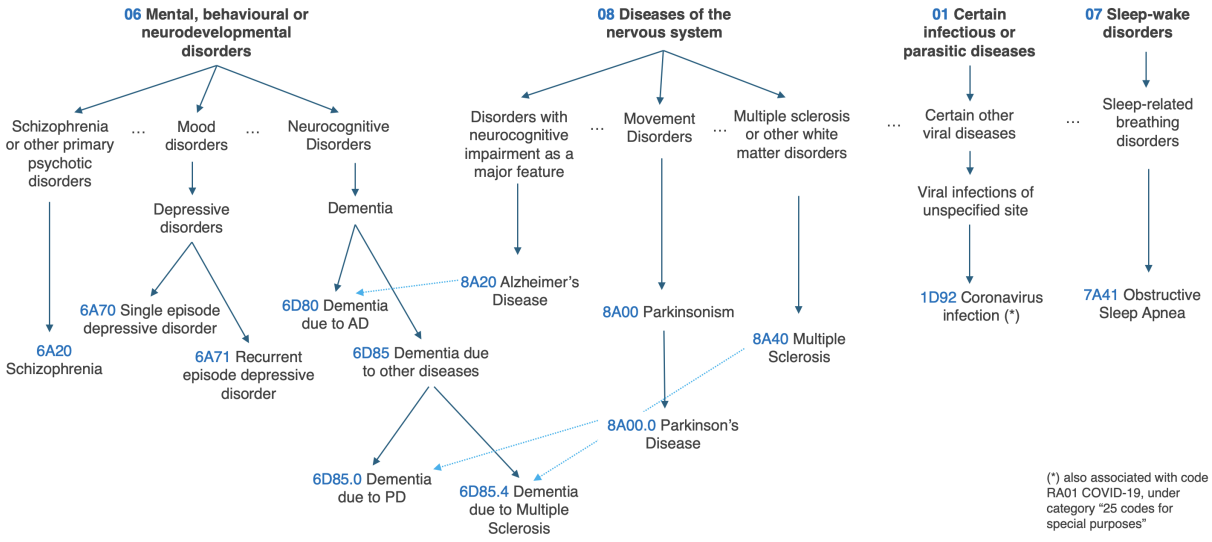


Figure A.1: Categorization of speech affecting diseases according to the International Classification of Diseases 11th Revision (ICD-11).

B

Appendix: In-the-wild data – is it suitable for disease detection?

In chapter 5, a pertinent concern was raised regarding the legitimacy of utilizing in-the-wild data for the detection of speech affecting diseases. On one hand, is the self reported health status a valid proxy for the true health status? On the other hand, can paralinguistic features and the emotional content be different when people are talking about their disease, as opposed to when they talk about book review, knitting, etc.? To address these questions, in a different work that focuses on the detection of depression and Parkinson's Disease from YouTube vlogs, we establish a comparison between in-the-wild data and publicly available benchmark datasets. These benchmark data sets were collected under controlled conditions in collaboration with medical healthcare professionals and are thus medically validated. Here, we briefly discuss those experiments and results because they support our claim that we can use in-the-wild data collected when no medically validated data are available.

B.1 Corpora

We used three corpora, described in detail in section 3.3: the WSM corpus, which contains controls, people suffering from PD and depressed subjects; the PC-GITA - standard dataset used in the speech community for the detection of Parkinson's disease, in Spanish; and the DAIC-WOZ - standard dataset used for the detection of depression, in English. In this work, with the goal of ensuring that the speech tasks recorded under controlled conditions and in-the-wild conditions data were similar, we only used a subset of exercises of the PC-GITA, including read words, read sentences, and spontaneous speech.

B.2 Experiments

We establish three classification baselines using distinct feature sets.

Baseline A.: we extracted eGeMAPS features, and perform the binary classification with a SVM.

Baseline B: we extracted i-vectors and performed the classification using PLDA classifier.

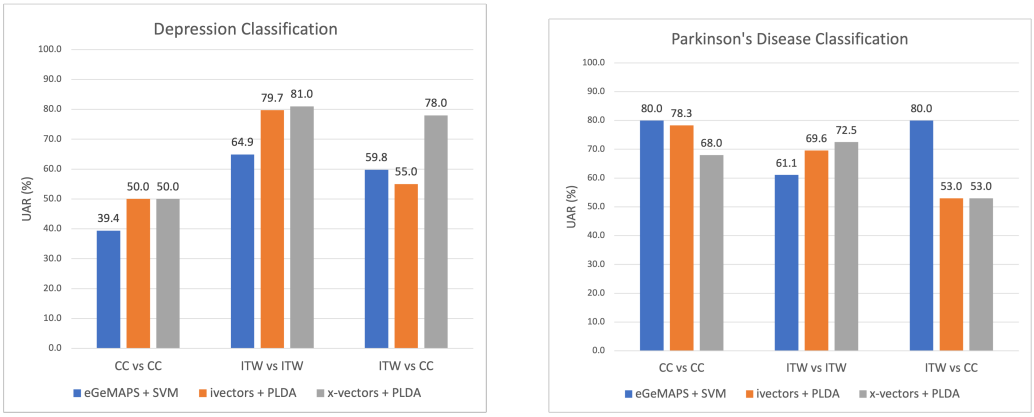
Baseline C: we extracted x-vectors and performed the classification also using PLDA.

Both baselines *B* and *C* follow the *Kaldi* recipe *egs/voxceleb/* [Povey et al., 2011].

These baseline experiments were designed for the release of the WSM corpus, aiming to employ standard methodologies. To facilitate comparison with small controlled condition (CC) datasets, we include a baseline system using SVM models combined with a general-purpose paralinguistic feature set, specifically eGeMAPS features. Additionally, we incorporate i-vector and x-vector based approaches using Probabilistic Linear Discriminant Analysis (PLDA). eGeMAPS, i-vectors and x-vectors were described in chapter 2. Studies conducted shortly before these experiments have demonstrated that i-vectors also contain information about the speaker's health status (e.g. [Hauptman et al., 2019]). Compared to i-vectors, x-vectors require shorter temporal segments to achieve optimal results and have been shown to be more robust to data variability and domain mismatches [Snyder et al., 2017a, 2018]. Similar to i-vectors, x-vectors have also been found to carry information about the speaker's health status (e.g. [Moro-Velazquez et al., 2020]).

We apply the three classification baselines to three different experiments:

1. *In-the-wild classification*, where we train and test the models using in-the-wild data. This experiment is abbreviated as *ITW vs ITW*;
2. *Controlled conditions classification*, where we train and test the models using the publicly available standard datasets. This experiment is abbreviated as *CC vs CC*;
3. *Cross-domain classification*, where we train the models with in-the-wild data, and test them with the data obtained in controlled conditions. This experiment is abbreviated as *ITW vs CC*. If a model trained with in-the-wild data and labels for self-reported health status, can perform close



(a) Depression **(b)** Parkinson's Disease

Figure B.1: Results obtained for the detection of depression (a) and Parkinson's disease (b). For each disease, we present the results for three sets of experiments: in controlled conditions using standard datasets (CC vs CC); in-the-wild (ITW vs ITW); and in cross domain experiments, where we trained using in-the-wild data and tested on controlled conditions datasets (ITW vs CC).

to, or better than the baseline models for standard datasets recorded in controlled conditions, we can interpret that as evidence to support our hypothesis that the self reported health status can be used as a proxy for true health status.

B.3 Results and discussion

Figure B.1 shows the classification results, in terms of UAR. Regarding the baselines for depression, using the DAIC-WOZ dataset (controlled conditions), we can observe that the performance of the three systems is chance level or worse. Hence, we conclude that detecting depression from speech alone in the DAIC-WOZ specifically is a difficult task. When using the WSM corpus (in-the-wild), the results are better, with the x-vector system being the best performing one. In the cross domain experiment, we can see that the performance of all the models improves when compared to training the models with the same domain, on controlled conditions.

Regarding the baselines for PD, using the PC-GITA dataset (controlled conditions), we were able to obtain, in the best case scenario, a UAR of 80.0%, with the eGeMAPS-based system. The x-vectors were the worst performing model. We hypothesise that this could be related to the mismatch between the language of the PC-GITA (Spanish data) and the language of the pre-trained x-vector extraction model (mostly English data). When using the WSM corpus (in-the-wild), we can generally notice a similar trend to the results obtained for depression, where the best performance was obtained by the baseline using x-vectors. This suggests that the baseline using x-vectors with PLDA can more robustly deal with variability of recording conditions obtained with in-the-wild data collection. In the cross domain experiment, when comparing to training the models with the same domain on controlled conditions, in

the best case scenario we are able to obtain the same performance in the two experiments, in the case where the modeling strategy is based on eGeMAPS. For the other two baseline systems, we observe a large performance drop. Again, we argue that the language mismatch in the cross domain experiment may hinder the performance of x-vectors and i-vectors. Nevertheless, given that for the model based on eGeMAPS, the performance remains the same across the two experiments, we consider that this result also supports our hypothesis that self-reported health status is a good proxy for true health status.

Both for depression and PD, we evaluated the statistical significance of the cross-domain experiments, when compared to the controlled conditions classification experiments. The statistical significance of these results was evaluated by paired t-tests that determine if the hypothesis that there is no statistically significant difference in the performance of equivalent models trained with data from CC or in-the-wild conditions can be rejected. We set the threshold of the probability necessary to reject the null hypothesis to 0.05. For the depression results, the improvement observed when training with in-the-wild data (figure B.1 (a) - ITW vs CC) compared to training with the standard dataset (figure B.1 (a) - CC vs CC) were found to be statistically significant. For the PD results, the difference in performance observed when training with in-the-wild data (figure B.1 (b) - ITW vs CC) compared to training with the standard dataset (figure B.1 (b) - CC vs CC) were found to not be statistically significant.

Overall, we conclude that these results support the claim that in-the-wild-data, in particular collected from YouTube vlogs is suitable for this task, and can be regarded as a starting point when no other data is available.

C

Appendix: multimodal OSA detection – supplementary material

This appendix provides details on the neural network architectures employed for OSA classification experiments. The hyperparameters for training were defined experimentally. All NNs, implemented in Pytorch [Paszke et al., 2019], were trained with Adam optimizer, and use binary cross entropy as loss function.

C.1 Neural networks for OSA detection from speech

As described in section 5.3.2, the neural architecture depends on the input type: x-vectors and KB features represent each audio segment with a fixed size vector, and thus are fed to a fully connected feed forward neural network; PASE+ features have a dimension that depends on the duration of the audio input, and thus are fed to a 1D CNN followed by a statistical pooling layer. All three neural networks were trained using cross entropy loss, in which each class was weighted by the inverse of its relative frequency

in the training folds.

The architectures of the neural networks used consist of convolutional blocks or feedforward blocks, and one final output linear layer with 2 nodes, followed by a softmax activation layer. Each fully connected block consists of: (i) one linear layer with 32; (ii) one batch normalization layer; (iii) one ReLU activation layer; and (iv) one dropout layer. Each convolutional block contains (i) one 1D-convolutional layer, which performs the convolution through time, with 64 and 32 filters (first and second blocks, respectively), kernel size of 3, stride 1, and padding to keep the time dimension constant; (ii) one batch normalization layer; (iii) one leaky ReLU activation layer; and (iv) one dropout layer. We carried out three experiments with the architectures and parameters described below.

A. X-vectors Experiment: The *x-vectors* are fed to a fully connected feed forward NN, with 3 fully connected blocks before the output layer. The NN was trained for 10 epochs, with batch size 64 and learning rate 0.001. The dropout probability was set to 0.5.

B. PASE+ Experiment: The PASE+ embeddings were fed to a CNN, which consists of two convolutional blocks, one statistical pooling layer which summarizes the time dimension to a fixed size output, and one fully connected block before the output layer. The fully connected block contains a Leaky ReLU activation layer, instead of a standard ReLU. The learning rate resembles 0.0001, the batch size was set to 32, the dropout probability was set to 0.7, and the network trained for 10 epochs.

C. KB features Experiment: The KB features are fed to a fully connected feed forward NN, equal in architecture and hyperparameters to the one described in experiment A. The only difference is the dropout probability, which was set to 0.5.

C.2 Neural networks for OSA detection from facial images

The NN architectures used for OSA detection from facial images also depend on the input type: for raw facial images, we used a CNN, and for KB features, BIF and embeddings we used feed forward NNs, with four fully connected blocks. The networks architectures are composed of 2 convolutional blocks or three feed forward blocks, and one final output linear layer with 2 nodes, followed by a softmax activation layer. Each convolutional block contains (i) one 2D-convolutional layer, with 8 filters, kernel size of 3, stride 1, and no padding; (ii) one ReLU activation layer; (iii) one max pooling layer with kernel size 2, and stride 2; and (iv) one dropout layer. Each fully connected block consists of: (i) one linear layer; (ii) one batch normalization layer; (iii) one ReLU activation layer; and (iv) one dropout layer. We carried out six experiments with the architectures and parameters described below.

A. Facial images Experiment: In this experiment we fed the raw images, previously resized to (100, 100) to a CNN. The dropout probability was 0.5. The learning rate resembles 0.0001, the batch size was set

to 32, and the networks were trained for 20 epochs.

B. Facial images Experiment with local attention: Experiment A was repeated, adding an attention layer after the the second convolutional block. The attention layer consists of: (1) one 2D-convolutional layer, with 8 filters, a kernel size of 3, stride 1, and padding to keep the dimension of the input constant; (2) one Sigmoid activation layer. The output of the attention layer, the attention scores, are multiplied element-wise by the output of the second convolutional layer, before being fed to the output layer. The idea of local attention is introduced by the convolutional layer, which filters the inputs and selects which pixels to give more weight. The remaining hyperparameters were the same as described above. This experiment was designed to provide some explanation on which part of the image was considered more relevant by the network, for the classification task.

C. Facial images Experiment with global attention: Experiment A was also repeated, using global attention. The global attention layer follows the standard attention layer architecture: (1) one fully connected linear layer, with 4232 nodes - this number corresponds to the dimension of the flattened output of the last convolutional layer; (2) one hyperbolic tangent layer; (3) one softmax layer. Similarly to the previous experiment, the attention scores are multiplied element-wise by the output of the second convolutional layer, before being fed to the output layer. The idea of global attention is introduced by the fully connected layer. The remaining hyperparameters were the same as described above.

D. Knowledge-based features Experiment: We fed the KB features to a fully connected NN. All linear layers have 2048 nodes, the learning rate resembles 0.01, the dropout probability was set to 0, the batch size to 128, and each NN was trained for 20 epochs.

E. Bio-inspired features Experiment: We fed BIF vectors to a fully connected NN. The linear layers at each of the fully connected blocks have 2048, 512, and 128 nodes. The learning rate resembles 0.001, the dropout probability was set to 0.5, the batch size to 128, and each NN was trained for 5 epochs.

F. Facial embeddings Experiment: The embeddings were fed to a fully connected NN. The linear layers at each of the fully connected blocks have 128, 64, and 32 nodes. The learning rate resembles 0.001, the dropout probability was set to 0.5, the batch size to 128, and each NN was trained for 5 epochs.

C.3 Neural networks for OSA detection from visual speech

The network that used lip reading embeddings as input for OSA classification contains two convolutional blocks, one fully connected block and one output linear layer with 2 nodes, followed by a softmax activation layer. Each convolutional block contains (i) one 1D-convolutional layer, which performs the convolution through time, with 64 and 32 filters (first and second blocks, respectively), kernel size of 3, stride 1, and no padding; (ii) one batch normalization layer; (iii) one ReLU activation layer; (iv) one

max pooling layer with kernel size 2, and stride 2; and (v) one dropout layer. The fully connected block consists of (i) one linear layer with 32 nodes; (ii) one batch normalization layer; (iii) one leaky ReLU activation layer; and (iv) one dropout layer.

The learning rate resembles 0.0001, the batch size was set to 64, the NN was trained for 10 epochs, using cross entropy loss, in which each class was weighted by the inverse of it's relative frequency in the training folds. Dropout was set to 0.5.

C.4 Neural network for OSA detection with early fusion of the three modalities

This section details the neural network architecture and parameters used for early fusion of three modalities: speech, facial images and visual speech, in *experiment B: Early fusion NN experiment*.

In this experiment, we fed the three embeddings to a NN that follows the structure represented in figure 5.4 (left). The speech embeddings and facial embeddings were fed to a fully connected block, with the same structure as described in section C.2. The visual speech embeddings were fed to two convolutional blocks, with the same structure as described in section C.3, except the number of filters, that were 64 and 16, respectively. The outputs of these layers were then summed and fed to a new fully connected block, before the final output layer. The linear layers in all fully connected blocks have 64 nodes. The learning rate resembles 0.0001, the batch size was set to 32, the dropout probability was set to 0.5, and the network trained for 5 epochs.

D

Appendix: disease detection across datasets – supplementary material

This appendix provides details on the neural network architectures employed as feature extractors for COVID-19 classification experiments.

D.1 Deep neural networks for feature extraction in COVID-19 detection

TDNN-F embeddings

These emebddings were extracted using a reduced version of the TDNN-F based network [Povey et al., 2018], proposed for speaker recognition by [Villalba et al., 2020]. The network architecture is summarized in Table D.1. Each block, with the exception of the statistics pooling layer, corresponds to a TDNN, TDNN-F or dense layer, followed by a Leaky-ReLU activation, a batch normalization layer and a dropout

layer. Cough embeddings are 128-dimensional vectors obtained at the output of the final dense layer (layer block 7).

Table D.1: TDNN-F embedding network architecture.

Layer	Layer type	Ctx. 1	Ctx. 2	Size	Inner size
1	TDNN	t-2:t+2	-	512	-
2	TDNN-F	t-2,t	t,t+2	1024	256
3	TDNN-F	t	t	1024	256
4	TDNN-F	t-2,t	t,t+2	1024	256
5	Dense	t	-	2048	-
6	Stats. Pool.	full seq.	-	2×2048	-
7	Dense (embedding)	-	-	128	-

The TDNN-F network was implemented in Pytorch. Training and fine-tuning used the Adam optimizer with a weighted sum of the loss functions of the involved tasks. Classification tasks used binary cross-entropy loss, while the regression task used the mean squared error loss. In the second stage, each class was weighted with the inverse of its frequency in the training subset to address the unbalanced nature of this dataset.

The network was trained for 500 and 20 epochs with batch sizes of 16 and 32 for the first and second stages, respectively. In the first stage, learning rates of 0.0001 and 0.001 were used to train the embedding network and the age and gender classification layers, respectively. For the second stage, these learning rates were reduced to 0.00005 and 0.0001. For the remaining tasks in the second stage, a learning rate of 0.0005 was used. Moreover, the loss corresponding to each task was given a weight, found through hyperparameter search.

CNN embeddings

The VGGish model [Hershey et al., 2017] is an adaptation for audio classification of the VGG network [Simonyan and Zisserman, 2015]. It comprises four blocks, each with one or two convolutional layers followed by a pooling layer. The output of the last pooling layer is flattened and followed by two fully-connected layers and an output layer. This model was originally trained with 5.4M hours of YouTube data. In this work, we used a simplified version as shown in Table D.2. Layers 1 to 7 correspond to the original architecture, with pre-trained weights from the original model. The top-level fully-connected layers in the original model were here substituted by lower-dimensionality layer 9 to facilitate fine-tuning with limited data. Layer 8 flattens and reduces dimensionality.

This CNN architecture is used in this work in two different settings. In both cases, the generated embeddings are 256-dimensional vectors (output of layer 8). The first setting corresponds to using the pre-trained model as generic feature extractor, weights are directly loaded from the original model. The second setting corresponds to fine tuning the model for COVID-19 detection using a balanced subset

Table D.2: Architecture of the simplified VGGish

Layer	Layer type	Output shape
1	Conv2D	(96,64,64)
2	MaxPooling2D	(48,32,64)
3	Conv2D	(48,32,128)
4	MaxPooling2D	(24,16,128)
5-6	Conv2D (x2)	(24,16,256)
7	MaxPooling2D	(12,8,256)
8	GlobalAvg.Pooling2D	(256)
9	FullyConnected	(64)
10	FullyConnected	(1)

of the COUGHVID dataset. In this case, layers 9 and 10 are included on top to allow for fine-tuning for COVID-19 detection. The weights of these two layers are initialized randomly and the whole CNN is fine-tuned for 150 epochs using cross-entropy loss, the Adam optimizer with a learning rate of 10^{-5} and a batch size of 64.

PASE+ features

The PASE+ extractors, both trained on Librishpeech and COUGHVID, were trained for 150 epochs. We used a batch size of 64 with a learning rate of 0.0005 and 0.001 for the workers and the encoders, respectively. 256-dimensional feature vectors are extracted for each 10 ms frame.

E

Appendix: A framework for multidisease detection – supplementary material

This appendix provides supplementary material to chapter 7, including further details on neural network architecture and training hyperparameters, features, and results.

E.1 Ambiguous coreference chain

In chapter 7, the feature *ratio of ambiguous coreference chains* was excluded from the analysis because (1) the findings in section 7.4.1.D suggest it is not robust to dataset shifts, not even with stratified normalization, and (2) the confidence intervals on the upper bound for both genders and both ASR systems were larger than the RI itself, which indicates a poor confidence on the derived RI. Figure E.1 shows the distribution of the *ratio of ambiguous coreference chains* on the reference population. It is clear that

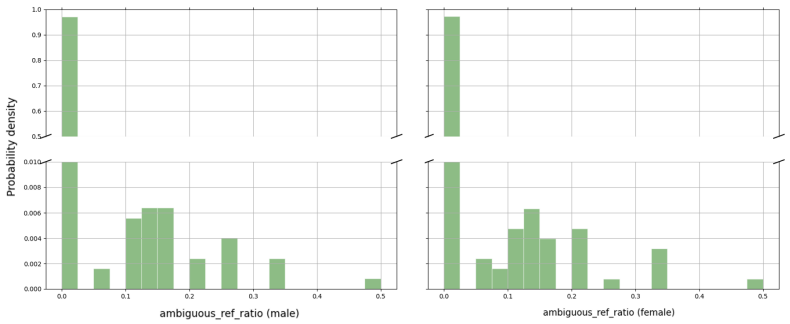


Figure E.1: Distribution of *ratio of ambiguous correference chains* on the reference population, based on whisper transcriptions.

Table E.1: Examples of picture descriptions in CLAC and the corresponding coreference chains identified by the coreference resolver.

ASR	Description	Coreference chains
wav2vec	"i see a mother not paying attention to what's happening in her kitchen she's drying the dishes and appears to be day dreaming while looking out the window because she's not paying attention the sink is overflowing causing a flood in the kitchen and her children are about to steal cookies from a cabinet well most likely about to get hurt because the stoolis about to tip over."	[Mother, Her, She'S, She'S, Her], [Kitchen, Kitchen]
whisper	"A young boy is walking down a path while attempting to fly a kite and his dog is following him. Behind him, there is a lake with a young girl on the beach, building a sand castle. On that same lake there is a gentleman on a dock, landing a fish. And on that lake out in the distance, there's a sailboat sailing. Meanwhile, in the foreground, there is a couple having a picnic. The woman is pouring a glass of wine. There's a stereo playing, and the man is reading a book. Down the street, there is a house with a car in the driveway and a tree in the front yard and a flag at Polstaff."	[Boy, His, Him, Him], [Lake, Lake, Lake]

the bulk of the distribution is very narrow as most samples correspond to zero. This resulted in a very narrow reference interval, and a very large confidence interval on the upper limit upon bootstrapping. Intuitively, one can understand that in a healthy population describing an image, there would rarely be any ambiguous pronouns, i.e., entities not explicitly mentioned or mentioned only cataphorically. Table E.1 shows examples of picture descriptions in CLAC and the corresponding coreference chains identified by the coreference resolver.

E.2 Hyperparameters for Neural Additive Models

The hyperparameter tuning for NAMs was performed with Bayesian optimization using Gaussian Processes, as implemented in scikit-optimize [Head et al., 2021], with 100 calls to the optimizer. The hyperparameters considered for tuning, with minor variations from those described by Agarwal et al. [2021], were as follows:

- learning rate: $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1\}$,
- dropout coefficient: $\{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$,
- weight decay: $[0.000001, 0.0001]$,

- feature dropout coefficient: $\{0, 0.05, 0.1, 0.2\}$
- output penalty coefficient: $[0.001, 0.1]$.

Additionally, the feature subnetworks were configured in one of the following ways: (i) one feedforwrad layer with 1024 hidden units, (ii) one feedforwrad layer with 512 hidden units, or (iii) three feedforward layers with 64, 64, and 32 hidden units. The activation functions for the hidden units were either ReLU or ExU, as introduced by Agarwal et al. [2021]. The batch size was set to 64.

Agarwal et al. [2021] suggested using an ensemble of 10-100 models for each NAM. In this work, given the 10-fold cross-validation setting, we defined each NAM as an ensemble of 3 models, which after cross-validation results in a total of 30 models. Future work may investigate increasing the number of models.

Table E.2 reports the hyperparameters that yielded the best performance for the classification of PD and AD.

Table E.2: Best parameters found for NAMs on classification of PD and AD, on PC-GITA and ADReSS, respectively. "Hidden units" shows the number of hidden layers as well as the number of neurons used in each layer for each feature network.

	PC-GITA	ADReSS
Learning rate	0.01	0.1
Dropout	0.8	0.6
Weight decay	1×10^{-6}	1×10^{-6}
Feature dropout	0.0	0.2
Output penalty	0.001	0.001
Num units	1024	1024
Activation	ExU	ExU

E.3 Classification results

The results, in terms of accuracy, for the complete set of experiments are presented in Table E.3, E.4, and E.5 for PC-GITA, ADReSS with whisper transcriptions, and ADReSS with wav2vec transcriptions, respectively.

Results for the NAM experiments are detailed in Table E.6.

Table E.3: Parkinson's disease classification results, using SVM and logistic regression, in terms of accuracy in [%]. *0M 1V* stands for 0-mean and unit variance normalization.

	SVM												LR								
	0 mean, 1 var				MinMax				No Norm				0 mean, 1 var				MinMax				No Norm		
CT	params	dev	test	MV	params	dev	test	MV	params	dev	test	MV	dev	test	MV	dev	test	MV	dev	test	MV
DT_{MSTD}																					
0.5	linear, C=1.0	58.7	60.7	64.0	poly, C=0.01, d=3	59.7	56.7	57.0	RBF, C=1.0	61.5	60.3	65.0	59.7	60.7	65.0	57.3	56.0	57.0	62.8	60.7	64.0
0.8	linear, C=1.0	63.9	64.0	67.0	linear, C=0.1	64.6	61.7	66.0	RBF, C=1.0	62.2	59.0	63.0	61.1	63.3	65.0	61.1	61.3	65.0	60.8	61.0	64.0
0.9	linear, C=1.0	63.5	63.7	67.0	linear, C=0.1	63.9	61.3	66.0	RBF, C=1.0	61.5	58.7	62.0	62.2	64.0	67.0	61.1	61.3	64.0	61.1	61.0	65.0
1.0	linear, C=0.1	63.2	61.3	64.0	linear, C=1.0	62.2	61.7	62.0	RBF, C=1.0	63.5	62.7	65.0	60.8	61.7	60.0	61.1	60.0	64.0	63.9	61.3	64.0
1.0	linear, C=0.1	63.2	61.3	64.0	linear, C=1.0	62.2	61.7	62.0	RBF, C=1.0	63.5	62.7	65.0	60.8	61.7	60.0	61.1	60.0	64.0	63.9	61.3	64.0
$DT_{MSTD-no-cap}$																					
0.5	RBF, C=0.1	59.7	57.3	59.0	RBF, C=1.0	62.8	61.0	65.0	RBF, C=1.0	57.3	57.0	60.0	59.4	58.7	62.0	61.8	59.7	64.0	55.6	55.0	59.0
0.8	linear, C=0.01	63.5	61.7	65.0	linear, C=0.01	66.0	66.0	67.0	RBF, C=0.1	59.4	56.3	56.0	58.7	58.0	57.0	62.2	64.3	68.0	55.6	57.3	62.0
0.9	linear, C=0.01	62.8	61.7	63.0	RBF, C=1.0	66.7	64.3	67.0	RBF, C=0.1	61.1	58.7	59.0	59.0	57.0	56.0	61.8	64.0	67.0	55.9	57.0	62.0
DT_{Q123}																					
0.5	linear, C=0.01	62.8	60.7	62.0	linear, C=0.01	66.3	66.7	72.0	RBF, C=1.0	56.3	55.7	57.0	67.0	67.3	71.0	67.0	66.3	69.0	64.2	62.7	65.0
0.8	linear, C=0.1	68.1	67.7	69.0	linear, C=0.01	69.1	69.3	76.0	RBF, C=1.0	55.9	56.0	57.0	69.4	68.3	71.0	69.1	69.3	75.0	64.2	63.3	67.0
0.9	linear, C=0.01	62.2	62.7	65.0	linear, C=0.01	69.4	69.7	77.0	RBF, C=1.0	56.3	56.0	57.0	69.1	67.7	70.0	69.8	68.3	72.0	64.6	62.3	65.0
1.0	linear, C=0.01	66.7	65.7	67.0	linear, C=0.01	69.1	69.3	75.0	RBF, C=1.0	57.3	56.3	57.0	68.1	66.0	71.0	64.9	66.7	74.0	67.7	68.0	73.0
DT_{RI}																					
0.5	poly, C=1, d=2	58.0	56.3	58.0	linear, C=1.0	68.8	68.3	73.0	RBF, C=1.0	54.2	53.7	54.0	56.3	57.7	61.0	70.5	69.7	74.0	54.5	54.0	54.0
0.8	RBF, C=0.1	59.0	57.0	58.0	linear, C=1.0	68.8	69.0	73.0	RBF, C=1.0	54.2	53.7	54.0	57.6	58.3	62.0	70.8	71.3	75.0	54.5	54.0	54.0
0.9	RBF, C=0.1	59.4	57.3	59.0	linear, C=0.1	67.4	69.0	71.0	RBF, C=1.0	54.2	53.7	54.0	57.6	58.7	62.0	70.1	70.0	72.0	54.5	54.0	54.0
1.0	RBF, C=1.0	60.4	59.7	64.0	linear, C=1.0	71.2	68.7	71.0	RBF, C=1.0	54.2	53.7	54.0	60.1	60.0	63.0	72.2	71.7	75.0	55.2	55.0	55.0
$DT_{Mahalanobis}$																					
0.5	poly, C=0.01, d=3	57.6	56.0	58.0	poly, C=1, d=2	65.6	61.3	62.0	linear, C=0.1	55.9	53.3	56.0	55.2	53.7	55.0	65.6	64.3	69.0	49.7	49.7	53.0
0.8	poly, C=1, d=2	62.5	61.7	65.0	linear, C=0.1	67.4	64.7	69.0	linear, C=0.1	55.6	55.3	58.0	60.8	61.7	66.0	66.7	65.0	70.0	53.5	52.3	51.0
0.9	linear, C=1.0	63.2	60.0	61.0	RBF, C=1.0	67.0	63.0	68.0	linear, C=0.01	56.3	55.3	56.0	62.2	62.7	66.0	67.4	64.0	69.0	54.5	54.7	53.0
1.0	RBF, C=0.1	62.8	60.7	65.0	linear, C=0.01	66.7	65.7	70.0	RBF, C=0.1	62.2	61.0	66.0	62.8	61.7	63.0	66.0	67.3	73.0	61.8	60.7	65.0
Features																					
0.5	linear, C=0.1	65.3	64.3	66.0	linear, C=1.0	65.6	63.3	67.0	poly, C=1, d=3	54.5	55.7	58.0	62.8	65.0	66.0	64.6	64.0	69.0	67.0	67.7	69.0
0.8	linear, C=0.1	68.8	67.0	70.0	linear, C=1.0	66.3	68.0	71.0	poly, C=1, d=3	52.1	53.7	54.0	67.0	67.0	69.0	64.2	65.3	70.0	67.0	67.7	69.0
0.9	linear, C=0.1	68.4	68.0	72.0	poly, C=1, d=2	67.7	67.3	73.0	poly, C=1, d=3	53.1	52.0	54.0	66.0	66.7	68.0	64.6	65.7	70.0	67.0	67.7	69.0
1.0	linear, C=0.1	67.0	67.3	68.0	poly, C=1, d=3	67.4	65.3	70.0	poly, C=1, d=3	53.8	55.3	58.0	64.6	66.3	68.0	66.3	66.3	71.0	64.9	66.3	69.0
Features regular norm																					
0.5	linear, C=0.1	65.6	65.0	67.0	poly, C=0.1, d=3	66.3	64.7	68.0	poly, C=1, d=3	54.5	55.7	58.0	64.9	66.0	67.0	63.9	63.3	66.0	67.0	67.7	69.0
0.8	linear, C=0.1	67.4	64.0	67.0	linear, C=1.0	66.0	64.3	68.0	poly, C=1, d=3	52.1	53.7	54.0	64.6	65.3	67.0	64.9	65.0	69.0	67.0	67.7	69.0
0.9	linear, C=0.1	66.7	64.0	66.0	linear, C=1.0	65.3	63.3	69.0	poly, C=1, d=3	53.1	52.0	54.0	64.2	65.3	67.0	64.6	64.0	68.0	67.0	67.7	69.0
1.0	linear, C=0.1	64.6	66.3	70.0	poly, C=1, d=2	65.3	65.3	67.0	poly, C=1, d=3	53.8	55.3	58.0	66.3	66.7	69.0	64.6	65.3	70.0	64.6	66.3	69.0

Table E.4: Alzheimer's disease classification results, using whisper transcriptions, using SVM and logistic regression, in terms of accuracy in [%]. *0M 1V* stands for 0-mean and unit variance normalization.

	SVM									LR					
	0M 1V			MinMax			No Norm			0M 1V		MinMax		No Norm	
CT	params	dev	test	params	dev	test	params	dev	test	dev	test	dev	test	dev	test
DT_{MSTD}															
0.5	poly, C=1, d=3	69.4	68.8	RBF, C=0.01	60.2	50.0	linear, C=0.01	67.6	45.8	63.9	60.4	57.4	66.7	63.9	64.6
0.7	linear, C=1.0	65.7	62.5	poly, C=0.1, d=2	60.2	62.5	poly, C=1, d=3	70.4	72.9	63.9	58.3	56.5	60.4	63.9	64.6
0.8	linear, C=1.0	65.7	60.4	RBF, C=0.01	60.2	50.0	poly, C=1, d=3	68.5	68.8	65.7	60.4	56.5	66.7	62.0	64.6
0.9	linear, C=1.0	65.7	62.5	RBF, C=1.0	61.1	64.6	poly, C=1, d=3	70.4	75.0	65.7	60.4	57.4	68.8	63.9	64.6
1.0	linear, C=1.0	66.7	58.3	RBF, C=1.0	63.0	60.4	poly, C=1, d=3	65.7	72.9	64.8	60.4	56.5	68.8	61.1	66.7
$DT_{MSTD-no-cap}$															
0.5	poly, C=0.1, d=3	65.7	66.7	poly, C=1, d=3	63.0	64.6	poly, C=1.0, d=2	74.1	75.0	61.1	62.5	57.4	75.0	66.7	64.6
0.7	poly, C=0.1, d=2	64.8	68.8	RBF, C=0.01	63.0	50.0	linear, C=0.1	73.1	66.7	59.3	64.6	58.3	68.8	66.7	66.7
0.8	linear, C=0.1	65.7	64.6	poly, C=0.1, d=3	64.8	60.4	linear, C=0.1	70.4	62.5	63.0	62.5	56.5	70.8	63.9	66.7
0.9	poly, C=0.1, d=2	64.8	64.6	poly, C=0.1, d=3	63.9	62.5	linear, C=0.1	70.4	68.8	60.2	64.6	58.3	70.8	62.0	70.8
1.0	RBF, C=0.01	65.7	50.0	linear, C=1.0	62.0	70.8	linear, C=0.1	70.4	66.7	60.2	62.5	58.3	72.9	64.8	66.7
DT_{Q123}															
0.5	RBF, C=0.01	60.2	50.0	RBF, C=1.0	63.0	66.7	linear, C=0.1	66.7	62.5	58.3	66.7	61.1	60.4	66.7	66.7
0.7	poly, C=1.0, d=2	61.1	64.6	RBF, C=1.0	64.8	62.5	linear, C=0.1	65.7	62.5	59.3	66.7	62.0	60.4	65.7	68.8
0.8	poly, C=1.0, d=2	62.0	68.8	RBF, C=1.0	66.7	64.6	linear, C=0.1	68.5	64.6	58.3	68.8	60.2	58.3	65.7	66.7
0.9	poly, C=1.0, d=2	61.1	70.8	RBF, C=1.0	64.8	64.6	linear, C=0.1	64.8	64.6	55.6	68.8	57.4	58.3	62.0	68.8
1.0	linear, C=1.0	63.0	68.8	RBF, C=1.0	67.6	62.5	linear, C=0.1	65.7	60.4	59.3	70.8	61.1	58.3	61.1	68.8
DT_{RI}															
0.5	RBF, C=1.0	57.4	72.9	linear, C=1.0	59.3	62.5	linear, C=1.0	65.7	70.8	61.1	60.4	56.5	68.8	68.5	72.9
0.7	linear, C=1.0	59.3	66.7	RBF, C=1.0	60.2	58.3	linear, C=1.0	66.7	75.0	63.9	60.4	56.5	62.5	70.4	77.1
0.8	RBF, C=0.01	59.3	50.0	RBF, C=1.0	62.0	60.4	linear, C=1.0	64.8	70.8	63.0	60.4	56.5	60.4	70.4	75.0
0.9	linear, C=1.0	58.3	66.7	RBF, C=1.0	63.0	62.5	linear, C=1.0	64.8	70.8	63.0	60.4	57.4	60.4	70.4	75.0
1.0	RBF, C=1.0	60.2	66.7	RBF, C=1.0	62.0	62.5	linear, C=1.0	65.7	70.8	59.3	64.6	55.6	60.4	68.5	75.0
$DT_{Mahalanobis}$															
0.5	linear, C=0.1	58.3	66.7	linear, C=0.01	54.6	58.3	linear, C=0.1	59.3	72.9	54.6	66.7	52.8	58.3	58.3	72.9
0.7	linear, C=1.0	54.6	62.5	RBF, C=1.0	56.5	58.3	RBF, C=1.0	52.8	50.0	52.8	62.5	59.3	62.5	38.9	60.4
0.8	linear, C=0.1	54.6	54.2	poly, C=0.1, d=2	55.6	56.3	RBF, C=1.0	55.6	52.1	53.7	54.2	55.6	56.3	38.9	62.5
0.9	RBF, C=0.01	53.7	50.0	RBF, C=1.0	56.5	58.3	RBF, C=0.01	50.0	50.0	53.7	54.2	56.5	56.3	38.9	39.6
1.0	linear, C=0.1	52.8	50.0	RBF, C=0.01	54.6	50.0	poly, C=0.01, d=2	46.3	50.0	54.6	50.0	55.6	62.5	41.7	58.3
Features															
0.5	linear, C=1.0	72.2	70.8	poly, C=1.0, d=2	69.4	72.9	RBF, C=0.01	57.4	50.0	71.3	66.7	68.5	64.6	68.5	62.5
0.7	linear, C=0.1	69.4	66.7	poly, C=1.0, d=2	70.4	70.8	RBF, C=0.01	56.5	50.0	67.6	66.7	65.7	62.5	62.0	64.6
0.8	linear, C=1.0	70.4	66.7	poly, C=1.0, d=2	71.3	72.9	RBF, C=0.01	56.5	50.0	65.7	66.7	65.7	66.7	63.9	64.6
0.9	linear, C=0.1	66.7	66.7	poly, C=1.0, d=2	69.4	70.8	RBF, C=0.01	56.5	50.0	67.6	64.6	63.0	66.7	63.0	66.7
1.0	linear, C=0.1	67.6	68.8	poly, C=1.0, d=2	75.9	68.8	RBF, C=0.01	58.3	50.0	65.7	64.6	65.7	66.7	63.0	64.6
Features regular norm															
0.5	linear, C=1.0	73.1	39.6	poly, C=0.1, d=3	72.2	50.0	RBF, C=0.01	57.4	50.0	70.4	39.6	67.6	50.0	68.5	62.5
0.7	linear, C=1.0	73.1	39.6	poly, C=1.0, d=2	74.1	50.0	RBF, C=0.01	56.5	50.0	69.4	39.6	68.5	50.0	62.0	64.6
0.8	linear, C=1.0	74.1	39.6	poly, C=1.0, d=2	73.1	50.0	RBF, C=0.01	56.5	50.0	68.5	39.6	65.7	50.0	63.0	64.6
0.9	linear, C=1.0	69.4	39.6	poly, C=1.0, d=2	71.3	50.0	RBF, C=0.01	56.5	50.0	67.6	39.6	65.7	50.0	63.9	66.7
1.0	linear, C=0.01	65.7	39.6	poly, C=0.1, d=3	70.4	50.0	RBF, C=0.01	58.3	50.0	68.5	39.6	63.0	50.0	64.8	66.7

Table E.5: Alzheimer's disease classification results, using wav2vec transcriptions, using SVM and logistic regression, in terms of accuracy in [%]. *0M 1V* stands for 0-mean and unit variance normalization.

	SVM									LR					
	0M 1V			MinMax			No Norm			0M 1V		MinMax		No Norm	
CT	params	dev	test	params	dev	test	params	dev	test	dev	test	dev	test	dev	test
DT_{MSTD}															
0.5	RBF, C=1.0	59.0	60.4	poly, C=1, d=3	54.0	58.3	poly, C=1, d=2	66.0	62.5	58.0	58.3	52.0	62.5	66.0	62.5
0.7	poly, C=1, d=2	60.0	64.6	poly, C=1, d=3	65.0	56.3	linear, C=1.0	67.0	58.3	56.0	54.2	48.0	58.3	65.0	58.3
0.8	RBF, C=1.0	63.0	58.3	poly, C=1, d=3	65.0	58.3	poly, C=1, d=2	69.0	56.3	59.0	58.3	53.0	62.5	62.0	58.3
0.9	RBF, C=1.0	63.0	56.3	poly, C=1, d=3	68.0	58.3	poly, C=1, d=2	70.0	58.3	60.0	58.3	53.0	62.5	62.0	56.3
1.0	RBF, C=1.0	59.0	60.4	poly, C=1, d=2	63.0	56.3	linear, C=1.0	69.0	52.1	59.0	58.3	47.0	56.3	64.0	56.3
$DT_{MSTD-no-cap}$															
0.5	linear, C=0.01	58.0	58.3	RBF, C=1.0	61.0	58.3	linear, C=0.01	69.0	60.4	53.0	56.3	51.0	70.8	63.0	54.2
0.7	poly, C=1, d=3	59.0	62.5	poly, C=1, d=3	58.0	64.6	poly, C=1, d=2	68.0	64.6	53.0	56.3	52.0	72.9	63.0	50.0
0.8	poly, C=1, d=2	60.0	60.4	linear, C=0.1	61.0	66.7	linear, C=0.01	68.0	60.4	54.0	62.5	52.0	70.8	60.0	50.0
0.9	poly, C=1, d=3	62.0	62.5	linear, C=0.01	58.0	68.8	linear, C=0.01	66.0	60.4	55.0	60.4	54.0	68.8	62.0	50.0
1.0	poly, C=0.1, d=3	60.0	62.5	poly, C=1, d=2	60.0	64.6	linear, C=0.01	65.0	60.4	51.0	58.3	50.0	58.3	58.0	47.9
DT_{Q123}															
0.5	poly, C=1, d=2	62.0	62.5	RBF, C=1.0	62.0	64.6	RBF, C=1.0	61.0	62.5	53.0	66.7	58.0	60.4	54.0	58.3
0.7	poly, C=1, d=2	63.0	64.6	RBF, C=1.0	67.0	70.8	linear, C=0.01	62.0	64.6	53.0	62.5	56.0	58.3	53.0	54.2
0.8	RBF, C=1.0	64.0	66.7	RBF, C=1.0	67.0	70.8	linear, C=0.01	63.0	64.6	51.0	66.7	57.0	62.5	52.0	54.2
0.9	RBF, C=1.0	65.0	66.7	RBF, C=1.0	67.0	70.8	linear, C=0.01	61.0	64.6	52.0	64.6	56.0	60.4	52.0	56.3
1.0	RBF, C=1.0	64.0	66.7	RBF, C=1.0	67.0	68.8	linear, C=0.01	60.0	64.6	55.0	64.6	55.0	58.3	54.0	62.5
DT_{RI}															
0.5	RBF, C=1.0	60.0	58.3	poly, C=1, d=2	59.0	62.5	linear, C=1.0	58.0	62.5	53.0	58.3	56.0	64.6	64.0	64.6
0.7	RBF, C=1.0	59.0	60.4	RBF, C=1.0	61.0	60.4	linear, C=1.0	58.0	64.6	54.0	58.3	56.0	64.6	64.0	64.6
0.8	RBF, C=1.0	59.0	62.5	linear, C=0.1	63.0	64.6	linear, C=1.0	59.0	62.5	54.0	58.3	55.0	62.5	66.0	64.6
0.9	RBF, C=1.0	59.0	62.5	RBF, C=1.0	63.0	62.5	linear, C=1.0	65.0	64.6	54.0	58.3	59.0	58.3	66.0	64.6
1.0	RBF, C=1.0	56.0	60.4	linear, C=0.1	63.0	58.3	linear, C=1.0	65.0	64.6	54.0	56.3	56.0	58.3	62.0	64.6
$DT_{Mahalanobis}$															
0.5	RBF, C=1.0	63.0	68.8	poly, C=0.01, d=3	58.0	62.5	RBF, C=0.1	63.0	50.0	60.0	68.8	55.0	68.8	65.0	64.6
0.7	RBF, C=1.0	56.0	54.2	linear, C=0.1	56.0	50.0	RBF, C=1.0	59.0	64.6	58.0	52.1	51.0	54.2	38.0	54.2
0.8	RBF, C=1.0	53.0	39.6	poly, C=0.1, d=3	59.0	54.2	RBF, C=1.0	60.0	64.6	52.0	47.9	57.0	43.8	37.0	54.2
0.9	RBF, C=1.0	53.0	35.4	poly, C=0.1, d=3	59.0	52.1	RBF, C=0.01	48.0	43.8	52.0	45.8	56.0	47.9	41.0	50.0
1.0	linear, C=0.1	53.0	45.8	poly, C=1, d=2	54.0	52.1	RBF, C=0.01	48.0	43.8	51.0	45.8	53.0	47.9	41.0	52.1
Features															
0.5	RBF, C=1.0	61.0	68.8	RBF, C=1.0	61.0	70.8	poly, C=0.01, d=3	48.0	43.8	56.0	62.5	61.0	66.7	65.0	58.3
0.7	RBF, C=1.0	61.0	66.7	RBF, C=1.0	62.0	70.8	poly, C=0.01, d=3	48.0	43.8	57.0	68.8	63.0	64.6	65.0	58.3
0.8	RBF, C=1.0	62.0	66.7	RBF, C=1.0	65.0	66.7	poly, C=0.01, d=3	48.0	43.8	55.0	64.6	59.0	62.5	66.0	58.3
0.9	poly, C=1, d=2	62.0	68.8	RBF, C=1.0	64.0	66.7	poly, C=0.01, d=3	48.0	43.8	56.0	60.4	59.0	62.5	64.0	62.5
1.0	RBF, C=1.0	60.0	70.8	poly, C=0.1, d=3	62.0	62.5	poly, C=0.01, d=3	48.0	43.8	56.0	62.5	61.0	62.5	62.0	54.2
Features regular norm															
0.5	linear, C=0.1	66.0	41.7	poly, C=1, d=3	63.0	43.8	poly, C=0.01, d=3	48.0	43.8	58.0	41.7	61.0	50.0	64.0	58.3
0.7	poly, C=1, d=2	60.0	43.8	poly, C=1, d=3	62.0	43.8	poly, C=0.01, d=3	48.0	43.8	56.0	41.7	59.0	50.0	65.0	58.3
0.8	linear, C=0.01	62.0	41.7	poly, C=1, d=3	62.0	43.8	poly, C=0.01, d=3	48.0	43.8	56.0	41.7	61.0	50.0	66.0	58.3
0.9	linear, C=0.01	61.0	41.7	RBF, C=1.0	61.0	43.8	poly, C=0.01, d=3	48.0	43.8	61.0	41.7	58.0	50.0	65.0	62.5
1.0	linear, C=0.1	60.0	41.7	RBF, C=1.0	61.0	43.8	poly, C=0.01, d=3	48.0	43.8	55.0	41.7	57.0	50.0	64.0	54.2

Table E.6: Classification results, using NAMs, in terms of accuracy (Acc), macro precision (P), macro recall (R), and macro F1, in [%].

				Dev Folds				Test				Test – Speaker MV			
CT	DT	norm	ASR	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Parkinson's Disease															
CT=1.0	DT_{RI}	MinMax	–	75.0	75.8	74.9	74.8	68.7	69.9	68.7	68.2	73.0	74.7	73.0	72.5
CT=1	DT_{Q123}	MinMax	–	72.2	73.0	72.2	72.0	66.3	67.3	66.3	65.9	68.0	69.1	68.0	67.5
CT=0.9	DT_{Q123}	MinMax	–	72.6	73.2	72.5	72.4	62.7	64.0	62.7	61.8	67.0	69.2	67.0	66.0
Alzheimer's Disease															
CT=1.0	Feats	MinMax	whisper	83.3	83.4	83.3	83.3	72.9	73.3	72.9	72.8	–	–	–	
CT=0.7	DT_{RI}	none	whisper	73.1	75.7	73.1	72.5	70.8	70.8	70.8	70.8	–	–	–	–
CT=0.5	DT_{RI}	MinMax	whisper	79.6	80.3	79.6	79.5	70.8	72.2	70.8	70.4	–	–	–	–
C=0.5	Feats	MinMax	whisper	84.3	84.4	84.3	84.2	75.0	75.0	75.0	75.0	–	–	–	–
CT=0.5	DT_{RI}	none	whisper	75.9	77.8	75.9	75.5	72.9	73.0	72.9	72.9	–	–	–	–
CT=1.0	Feats	none	whisper	81.5	81.9	81.5	81.4	72.9	73.0	72.9	72.9	–	–	–	–

F

Appendix: LLMs for AD detection – supplementary material

Figures F.1 to F.7 show the prompts used to query the LLMs. The examples in prompt strategy 4.1 are from DementiaBank corpus [Becker et al., 1994], but do not integrate the ADReSS corpus. These examples correspond to the manual transcriptions. Similar prompts were defined for the ASR generated transcriptions.

P1.1:

USER:

You are an expert evaluator that detects if a person suffers from Alzheimer's Disease, from their language.
Person A is describing the "Cookie Theft" image using spontaneous speech. Given the transcription of the person's description, your task is to predict if the person suffers from Alzheimer's Disease (AD).
Provide the output using the following json format:

{'comments': step-by-step explanation, with maximum 300 tokens.
'alzheimers_prediction': YES/NO,
'confidence_in_prediction': high/low}

No other output should be provided.

DESCRIPTION: *<picture_description>*

Figure F.1: Prompt strategy P1.1.

Figure F.8 presents the suboptimal results obtained with *Llama-2-13B*. We have also performed

P1.2:

USER:

You are an expert evaluator that detects if a person suffers from Alzheimer's Disease, from their language.
Person A is describing the "Cookie Theft" image using spontaneous speech. Given the transcription of the person's description, your task is to predict if the person suffers from Alzheimer's Disease (AD). Notice that the speech of a person suffering from AD is characterized by word-finding difficulties, repetitions, reduced vocabulary, an overuse of indefinite and vague terms, and inappropriate use of pronouns. Furthermore, the discourse of AD patients is described as fluent but not informative, characterized by incomplete and short sentences, and lacking coherence and cohesion.
Provide the output using the following json format:

{'comments': step-by-step explanation, with maximum 300 tokens.
'alzheimers_prediction': YES/NO,
'confidence_in_prediction': high/low}

No other output should be provided.

DESCRIPTION: *<picture_description>*

Figure F.2: Prompt strategy P1.2.

P1.3:

USER:

You are an expert evaluator that detects if a person suffers from Alzheimer's Disease, from their language.
Person A is describing the "Cookie Theft" image using spontaneous speech. Given the transcription of the person's description, your task is to predict if the person suffers from Alzheimer's Disease (AD). Notice that the speech of a person suffering from AD is characterized by word-finding difficulties, repetitions, reduced vocabulary, an overuse of indefinite and vague terms, and inappropriate use of pronouns. Furthermore, the discourse of AD patients is described as fluent but not informative, characterized by incomplete and short sentences, and lacking coherence and cohesion.
The image can be described using seven concepts (woman doing dishes, sink overflowing, boy on stool, children stealing cookies, girl reaching for cookie, stool falling, woman not noticing). For example, with reference to the first concept, "the mother is drying a plate" or "lady do dishes" is considered "accurate and complete; "lady with dishes" or "the mother is standing by the sink" is considered accurate but incomplete; "the woman is washing clothes" is considered inaccurate.
Provide the output using the following json format:

{'comments': step-by-step explanation, with maximum 300 tokens.
'alzheimers_prediction': YES/NO,
'confidence_in_prediction': high/low}

No other output should be provided.

DESCRIPTION: *<picture_description>*

Figure F.3: Prompt strategy P1.3.

P1.4:

USER:

You are an expert evaluator that detects if a person suffers from Alzheimer's Disease, from their language.
Person A is describing the "Cookie Theft" image using spontaneous speech. Given the transcription of the person's description, your task is to predict if the person suffers from Alzheimer's Disease (AD). Notice that the speech of a person suffering from AD is characterized by word-finding difficulties, repetitions, reduced vocabulary, an overuse of indefinite and vague terms, and inappropriate use of pronouns. Furthermore, the discourse of AD patients is described as fluent but not informative, characterized by incomplete and short sentences, and lacking coherence and cohesion.
The image can be described using seven concepts (woman doing dishes, sink overflowing, boy on stool, children stealing cookies, girl reaching for cookie, stool falling, woman not noticing). For example, with reference to the first concept, "the mother is drying a plate" or "lady do dishes" is considered "accurate and complete; "lady with dishes" or "the mother is standing by the sink" is considered accurate but incomplete; "the woman is washing clothes" is considered inaccurate.
Provide the output using the following json format:

{'comments': step-by-step explanation, with maximum 300 tokens.
'alzheimers_prediction': YES/NO,
'confidence_in_prediction': high/low}

No other output should be provided.

DESCRIPTION: www . the boy's getting cookies out of the cookie jar . he's handing one to a girl . the lad the stool he's standing on is falling . the lady's drying dishes . the sink is running over . the water's turned on full . um cups are sitting on the counter, plates sitting on the counter . puddle of water's on the floor . little girl is saying . shh don't tell anybody . and the cookie jar looks like it's ready to fall out . and the cookie jar is full, clear full . that's about all I see that's going on .

ASSISTANT:

{'comments': 'The description seems to be complete, mentioning the main entities in the picture: the kitchen, the girl, the boy, the lady or mother, the cookie jar and the overflowing sink. The description is also coherent. lexically diverse, and does not seem to evidence wording finding difficulties.',
'alzheimers_prediction': NO ,
'confidence_in_prediction': high }

USER:

DESCRIPTION: mm dishes are being dried . and the child is getting some cookies out of the jar . the the uh plant stand or the stand he's on is looks as though it's crooked . and the water's going over in the sink from the sink . is that all ? and then the boy was getting the cookies out of the jar and the cover is off . but she he's giving her a cookie and the stool is turning over and the water in the sink is boiling over or flowing over . and she's drying dishes .

ASSISTANT:

{'comments': 'The description seems confusing, without clearly identifying the subjects, and simply mentioning "she". The description has a poor lexically diversity. It seems to evidence wording finding difficulties, eg. "the the uh plant".',
'alzheimers_prediction': YES ,
'confidence_in_prediction': high }

USER:

DESCRIPTION: <picture_description>

Figure F.4: Prompt strategy P1.4.

P1.5:

USER:

You are an expert fluency evaluator. Your task is to evaluate the description of an image provided in spontaneous speech by a person. The evaluation should focus on identifying word finding difficulties, repetitions, reduced vocabulary, an overuse of indefinite and vague terms, inappropriate use of pronouns and lack of coherence and cohesion.
The image can be described using seven concepts (woman doing dishes, sink overflowing, boy on stool, children stealing cookies, girl reaching for cookie, stool falling, woman not noticing). For example, with reference to the first concept, "the mother is drying a plate" or "lady do dishes" is considered "accurate and complete; "lady with dishes" or "the mother is standing by the sink" is considered accurate but incomplete; "the woman is washing clothes" is considered inaccurate.
Provide the output using the following json format:

{'comments': step-by-step explanation, with maximum 300 tokens.
'issues': YES/NO,
'confidence': high/low}

No other output should be provided.

DESCRIPTION: *<picture_description>*

Figure F.5: Prompt strategy P1.5.

P2.1:

USER:

You are an expert fluency evaluator.
Person A is describing the "Cookie Theft" image using spontaneous speech. The image can be described using seven concepts (woman doing dishes, sink overflowing, boy on stool, children stealing cookies, girl reaching for cookie, stool falling, woman not noticing). Given the transcription of the person's description, your task is to evaluate the text in terms of coherence, lexical diversity, sentence length and word finding difficulties, using scores between 0 and 1.
Provide the ratings in a json format such as the example below. No other outputs.

{'text_coherence': number between 0 and 1,
'lexical_diversity': number between 0 and 1,
'sentence_length': number between 0 and 1,
'word_finding_difficulties': number between 0 and 1}

DESCRIPTION: *<picture_description>*

Figure F.6: Prompt strategy P2.1.

P2.2:

USER:

You are an expert fluency evaluator, that works in the medical domain to support medical screening.
Person A is describing the "Cookie Theft" image using spontaneous speech. The image can be described using seven concepts (woman doing dishes, sink overflowing, boy on stool, children stealing cookies, girl reaching for cookie, stool falling, woman not noticing). Given the transcription of the person's description, your task is to evaluate the text in terms of coherence, lexical diversity, sentence length and word finding difficulties, using scores between 0 and 1. Then you evaluate if the person is likely to suffer from Alzheimer's Disease.
Provide the ratings in a json format such as the example below. No other outputs.

{'text_coherence': number between 0 and 1,
'lexical_diversity': number between 0 and 1,
'sentence_length': number between 0 and 1,
'word_finding_difficulties': number between 0 and 1,
'alzheimers_prediction': YES/NO,
'confidence_in_prediction': high/low}

DESCRIPTION: *<picture_description>*

Figure F.7: Prompt strategy P2.2.

preliminary experiments with *Llama-2-7B*, but the output frequently failed to comply with the requested format, thus not allowing the automatic analysis of the results.

Table F.1 reports the results obtained for task 2, using *Llama-2-13B* for the five classifiers – SVM, LDA, 1NN, DT, and RF – that were used in the ADReSS baseline Luz et al. [2020].

	Llama 13B		
	#Fail	Acc_{train}	Acc_{test}
Manual transcriptions			
P1.1	0/0	56.5	54.2
P1.2	0/0	48.1	56.3
P1.3	0/0	55.6	54.2
P1.4	4/1	50.9	47.9
P1.5	18/5	51.9	56.3
P2.2	0/0	50.0	50.0
Whisper transcriptions			
P1.1	0/0	55.6	50.0
P1.2	0/0	49.1	54.2
P1.3	0/0	54.6	52.1
P1.4	1/1	53.7	52.1
P1.5	22/11	53.7	50.0
P2.2	0/0	50.0	50.0
Wav2vec transcriptions			
P1.1	0/0	45.4	33.3
P1.2	0/0	49.1	39.6
P1.3	0/0	52.8	47.9
P1.4	0/0	50.9	47.9
P1.5	3/2	50.0	47.9
P2.2	0/0	50.0	50.0
Mean	3/1	51.5	49.7

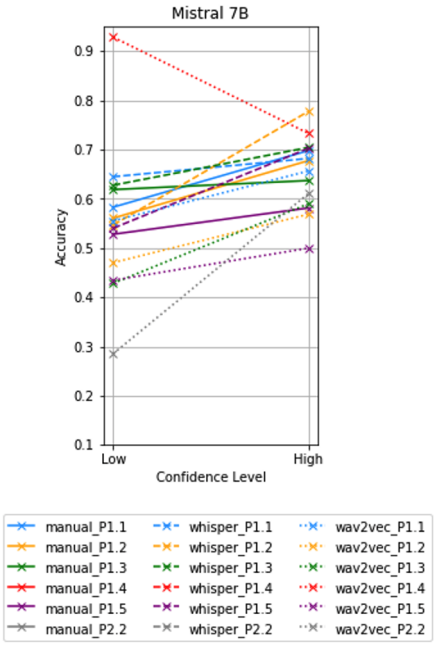


Figure F.8: Task 1 results for *Llama-2-13B*. The table on the left shows AD classification accuracy in %. *#Fail* denotes the number of examples for which the model failed to follow the output instruction (identified on train/test). The figure on the right shows the combined train and test accuracy per confidence level. A minimum of 10 instances per confidence level is required for the prompting strategy to be plotted.

Table F.1: AD classification based on macro-descriptors.

		Llama2 13B				
		SVM	LDA	1NN	DT	RF
Train set: 10-Fold CV						
Manual	P2.1	65.7	66.7	50.9	67.6	65.7
	P2.2	67.6	66.7	54.6	67.6	67.6
Whisper	P2.1	68.5	68.5	63.9	66.7	67.6
	P2.2	64.8	65.7	56.5	69.4	69.4
Wav2vec	P2.1	62.0	60.2	52.8	63.0	63.0
	P2.2	57.4	57.4	57.4	60.2	59.3
Mean				63.1		
Mean (ASR)				62.7		
Test set						
Manual	P2.1	68.8	72.9	54.2	68.8	68.8
	P2.2	72.9	70.8	68.8	72.9	72.9
Whisper	P2.1	66.7	62.5	64.6	58.3	58.3
	P2.2	54.2	64.6	54.2	64.6	64.6
Wav2vec	P2.1	64.6	66.7	58.3	66.7	66.7
	P2.2	64.6	64.6	41.7	52.1	64.6
Mean				63.8		
Mean (ASR)				61.1		