Thèse n° 9785

# Novel Methods For Detection And Analysis Of Atypical Aspects In Speech

Présentée le 26 mai 2023

Faculté des sciences et techniques de l'ingénieur
Laboratoire de l'IDIAP
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Julian David FRITSCH

Acceptée sur proposition du jury

Prof. J.-Ph. Thiran, président du jury
Dr J.-M. Odobez, Dr M. Magimai Doss, directeurs de thèse
Prof. J. R. Orozco Arroyave, rapporteur
Dr F. Ringeval, rapporteur
Dr D. Thanou, rapporteuse

École
polytechnique
fédérale
de Lausanne

2023

# Acknowledgements

This thesis is the result of the support of many people. First and foremost, I thank my supervisor Dr. Mathew Magimai Doss, who advised me, taught me many things, and needed lots of patience in doing so. Likewise, I thank my thesis director Dr. Jean-Marc Odobez. My thesis committee provided insightful comments on my thesis. I, therefore, thank Prof. Thiran, Dr. Thanou, Prof. Orozco Arroyave, and Prof. F. Ringeval for taking the time to read it.

This thesis contains work done in collaboration. Therefore I would like to thank: Pavankumar Dubagunta for his help and guidance, Bence Halpern, Camilo Vazquez-Correa, and Esaú Villatoro Tello. In my Ph.D., I met many peers through my funding project, TAPAS, an EU Horizon 2020 Marie Sklodowska-Curie project (grant agreement No 766287). This experience turned out a great opportunity, despite the two-year COVID-19 break. Especially I thank Prof. Elmar Nöth for referring me to Mathew for this project.

Idiap has made my Ph.D. specifically pleasurable. I am very grateful to the secretariat for among other things providing me with an amazing place to stay and the ease of filing travel reimbursements. Many colleagues made my time at Idiap an amazing time: Pavan, Dhananjay, Lesly, Enno, Apoorv, François, Florian, Banri, Bastian, Angel, Weipeng, Suraj, Angelos, Parvaneh, Suhan, Pablo, Neha, Zohreh, Amir, Tilak, and Eklavya and many more. At last, I thank my family and Laurie for their encouragement and support.

*Lausanne, February 15, 2023* Julian

# Abstract

Atypical aspects in speech concern speech that deviates from what is commonly considered normal or healthy. In this thesis, we propose novel methods for detection and analysis of these aspects, e.g. to monitor the temporary state of a speaker, diseases that manifest in speech, or people that have trouble producing speech. To overcome data scarcity, most methods in this thesis depend on auxiliary resources; to comply with clinicians, prior knowledge and explainability are taken into account.

In the first part of this thesis, we augment methods that aim to directly assess atypical speech with convolutional neural networks (CNN). With the goal of inducing prior knowledge about atypical speech into CNNs, we present findings in the context of Alzheimer's disease detection and severity estimation: We demonstrate that filtering the waveforms to focus on voice-source-related frequencies and increasing the input segment length to capture prosody has beneficial effects. Additionally, we explore incorporating phonetic knowledge into CNNs: By using CNN-based models trained for articulation prediction that are fine-tuned on continuous sleepiness estimation. Furthermore, we propose methods for detecting and estimating breathing impairments in people with Parkinson's disease. We compare hand-crafted features that model voice-source information and embeddings extracted from CNNs and find they are well-suited.

The second part of this thesis presents a novel method for intelligibility assessment of people with dysarthria. Intelligibility is a clinical measure of the severity of dysarthria. Typically assessed as an aggregate over a set of utterances by a speaker, we emulate the subjective listening tests by performing utterance verification using phonetic features on all of a speaker's utterances, aggregate them into the speaker's intelligibility score, and demonstrate this scheme's robustness through several variations. The same scheme was applied to emulate a human listening test, where listeners had to differentiate between before and after lip filler surgery. The intelligibility assessment scheme is extended into pronunciation feedback: Expected pronunciation is modeled by training one hidden Markov model per phoneme on healthy speech. Given a prompt and its corresponding dysarthric utterance, we can estimate by how much a phoneme deviates from its expected pronunciation and give a phoneme-level assessment.

**Keywords:** convolutional neural networks, articulatory features, Alzheimer's disease, degree of sleepiness, Parkinson's disease, speech intelligibility, dysarthria

# Résumé

Les aspects atypiques de la parole signifient la parole qui s'écarte de ce qui est considéré comme normal ou sain. Dans cette thèse, nous proposons de nouvelles méthodes de détection et d'analyse de ces aspects, par exemple pour surveiller l'état temporaire d'un locuteur, les maladies qui se manifestent dans la parole ou les personnes qui ont du mal à produire la parole. Pour maitriser la pénurie des données, la plupart des méthodes de thèse dépendent de ressources auxiliaires ; pour se conformer aux cliniciens, les intuitions humaines et des méthodes interprétables sont prises en compte.

Dans la première partie de cette thèse, nous augmentons les méthodes qui visent à évaluer directement la parole atypique avec des réseaux de neurones convolutifs (CNN). Dans le but d'induire des connaissances préalables sur le discours atypique dans les CNN, nous présentons des résultats dans le contexte de la détection de la maladie d'Alzheimer et de l'estimation de la gravité : nous montrons que filtrer des signaux pour se concentrer sur les fréquences liées à la source vocale et augmenter la longeur de segment d'entrée pour capturer la prosodie ont des effets bénéfiques. De plus, nous explorons l'intégration des connaissances phonétiques dans les CNN : en utilisant des modèles basés sur des CNN entraînés pour le prédicteur d'articulation, qui sont affinés sur l'estimation continue de la somnolence. De plus, nous proposons des méthodes de détection et d'estimation des troubles respiratoires des personnes atteintes de la maladie de Parkinson. Nous comparons les fonctionnalités artisanales qui modélisent les informations de source vocale et les intégrations extraites des CNN et constatons qu'elles sont bien adaptées.

La deuxième partie de cette thèse présente une nouvelle méthode d'évaluation de l'intelligibilité des personnes atteintes de dysarthrie. L'intelligibilité est une mesure clinique de la sévérité de la dysarthrie. Généralement évalués comme un agrégat sur un ensemble d'énoncés par un locuteur, nous émulons les tests d'écoute subjectifs en effectuant une vérification d'énoncé sur tous les énoncés d'un locuteur, les agrégeons dans le score d'intelligibilité de l'orateur et démontrons la robustesse de ce schéma à travers plusieurs variations. Le même schéma a été appliqué pour émuler un test d'écoute humaine, où les auditeurs devaient faire la différence entre avant et après la chirurgie de remplissage des lèvres. Le schéma d'évaluation de l'intelligibilité est étendu à un retour de prononciation : la prononciation attendue est modélisée en entraînant un modèle de Markov caché par phonème sur une parole saine. Nous pouvons estimer de combien un phonème s'écarte de sa prononciation attendue et donner une évaluation au niveau du phonème.

**Mots-clés :** réseaux de neurones convolutifs, caractéristiques articulatoires, maladie d'Alzheimer, degré de somnolence, maladie de Parkinson, intelligibilité de la parole, dysarthrie

# Zusammenfassung

Atypische Aspekte in Sprache meinen Sprache, die von dem abweicht, was allgemein als normal oder gesund angesehen wird. In dieser Doktorarbeit schlagen wir neue Methoden zur Detektion und Analyse dieser Aspekte vor, z.B. den vorübergehenden Zustand eines Sprechers, Krankheiten, die sich in der Sprache manifestieren, oder Menschen die Schwierigkeiten beim Sprechen haben. Um Datenknappheit zu kompensieren, sind die meisten Methoden in dieser Arbeit auf Hilfsressourcen angewiesen; Um den Klinikern zu helfen, werden menschliche Intuitionen und Erklärbarkeit der Methoden berücksichtigt.

Im ersten Teil dieser Arbeit ergänzen wir Methoden, die darauf abzielen, atypische Sprache direkt mit Convolutional Neural Networks (CNN) zu erfassen. Mit dem Ziel, Wissen über atypische Sprache in CNNs zu induzieren, präsentieren wir Ergebnisse im Zusammenhang mit der Erkennung und Schweregradschätzung von Alzheimer: Wir zeigen, dass das Filtern der Signale den Fokus auf stimmquellenbezogene Frequenzen legt und längere Eingangssegmente erleichtern dem Modell Prosodie zu erkennen, was positive Auswirkungen auf die Ergebnisse hat. Darüber hinaus untersuchen wir die Einbeziehung von phonetischem Wissen in CNNs: Durch die Verwendung von CNN-basierten Modellen, die für Artikulationsprädiktor trainiert wurden und auf kontinuierliche Schläfrigkeitsschätzung trainiert werden. Des Weiteren schlagen wir Methoden zur Erkennung und Einschätzung von Atembeeinträchtigungen bei Menschen mit Parkinson vor. Wir vergleichen Merkmale, die extrahierte Sprachquelleninformationen modellieren mit Merkmalen aus CNNs, und finden, dass beide gut geeignet sind.

Der zweite Teil dieser Arbeit stellt eine neue Methode zur Bewertung der Verständlichkeit von Menschen mit Dysarthrie vor. Die Verständlichkeit ist ein klinisches Maß für den Schweregrad der Dysarthrie. Typischerweise als Aggregat über eine Reihe von Äußerungen eines Sprechers bewertet, emulieren wir die subjektiven Hörtests, indem wir alle Äußerungen eines Sprechers einzeln Verifizieren, und das in eine Verständlichkeitsbewertung des Sprechers aggregieren. Wir zeigen die Robustheit dieses Schemas durch mehrere Variationen. Das gleiche Schema wurde angewendet, um einen menschlichen Hörtest zu emulieren, bei dem die Zuhörer zwischen vor und nach einer Lippenfüller-Operation unterscheiden mussten. Die Methode zur Bewertung von Verständlichkeit wird zu einem Sprech-Feedback erweitert: Die erwartete Aussprache wird modelliert, indem ein Hidden-Markov-Modell pro Phonem auf gesunder Sprache trainiert wird. Zwischen einem Prompt und der dysarthrischen Äußerung können wir abschätzen, um wieviel ein Phonem von seiner erwarteten Aussprache abweicht, und eine

Bewertung auf Phonemebene abgeben.

**Stichwörter:** Convolutional Neural Networks, Artikulationsmerkmale, Alzheimer, Parkinson, Sprachverständlichkeit, Hidden Markov Modell, Dysarthrie

# Contents

# List of Figures

# List of Tables

# 1 Introduction

Atypical aspects in speech concern speech that deviates from what people consider healthy normal speech. Depending on the strength of atypical aspects, they may reduce how intelligible speech is to the communication partner. Atypical aspects can stem from short-term speaker states, from difficulties with language or from any part of our complex speech-production process. The human speech signal can be used as a biological signal that can be translated to biomarkers reflecting the phenotype of certain diseases, that manifest in speech. Estimating these biomarkers offers the possibility of early detection, monitoring, and better treatment of diseases. Compared to other biomarkers, capturing speech signals is cheap and scalable and therefore has a large potential. Population screening for diseases is already part of many healthcare systems (e.g. cancer screening); detection of atypical aspects in speech can complement these screenings. Successful screening methods will save the healthcare system money, and prolong lives, since there are diseases for which no cure exists, such as Alzheimer's disease, Parkinson's disease, or cancer. The ideal scenario is illustrated in Figure 1.1 (generated with Stable Diffusion[1]): Screening and therapy should be available from the comfort of our living room.

## 1.1 Motivation

Atypical aspects are often highly individual and vary even from day to day. This makes it harder for human raters such as clinicians to rate/grade them. The medical context further comes with particular challenges, such as respecting the ethics and privacy of speakers. Automatic methods offer the potential to personalize speech analysis. In this thesis, the following challenges are addressed:

1. Data scarcity, since speech is subject to data protection regulation, and in a clinical field requires a physician's or phonetician's labeling, making it expensive. Self-reporting error is an alternative, but it might be error-prone and not accepted by clinicians.

[1] https://github.com/CompVis/stable-diffusion

Figure 1.1: Screening scenario of "an older person doing an exercise on a tablet sitting in her cozy living room" created with Stable Diffusion.

2. The definition of atypical speech and how it relates to normal speech. Labels of atypical speech based on a clinician's opinion. Normal speech is usually collected from speakers without any known atypicality. However, speech is highly individual, making the definition of an expected pronunciation or speaking style challenging.

3. Explainability is expected from methods in a clinical scenario so that physicians and clinicians trust them. Likewise, end-users need to understand the results of the analysis.

The goal of this thesis is to develop novel methods for the detection and analysis of atypical aspects in speech that address these challenges.

## 1.2 Contributions

Regarding the above-mentioned challenges, this thesis contributes to the following research questions:

1. We explore methods to induce human knowledge and intuitions about diseases into raw waveform CNN models. For Alzheimer's detection, we show that CNNs benefit from zero frequency filtering the signals before feeding to the network (Cummins et al., 2020). Further, we discovered that CNNs benefit from longer input segments, experimentally, we find 4 seconds to be optimal (Villatoro-Tello et al., 2021). To overcome data scarcity, we pre-trained CNNs to predict articulatory features, then fine-tuned them to predict sleepiness in speech, which yielded improvements (Fritsch et al., 2020).

2. For breathing impairment detection in Parkinson's disease patients, we find that statistics of features that describe the vocal vibrations as well as statistics of CNN-based embeddings work well (Vásquez-Correa et al., 2021).

3. For dysarthric intelligibility estimation, we propose a new scheme that emulates a human listening test by comparing dysarthric utterances to healthy reference and estimating the deviation in a phonological posterior space (Fritsch and Magimai-Doss, 2021). This concept also verifies the results of an A-B listening of recordings from before and after lip-filler surgery and evaluates synthetic pathological speech (Halpern et al., 2021).

4. We propose an HMM-based method for modeling healthy speech to detect pronunciation errors in speakers with dysarthria. The method allows a phoneme-based analysis at accuracies above 80%.

## 1.3 Thesis outline

Figure 1.2 shows a schematic overview of this thesis. Below, the thesis organization is briefly described.

Chapter 2 provides a background to this thesis. It introduces some re-occurring terminology, overviews speech production, different types of atypical speech, and overviews atypical speech analysis literature.

Chapter 3 describes our proposed methods to induce prior knowledge and intuitions about diseases into raw waveform CNN models. We present findings on Alzheimer's disease and sleepiness estimation.

Chapter 4 presents a comparison of methods to detect breathing impairment in Parkinson's disease patients. It describes a study of different types of features.

Chapter 5 presents novel methods to estimate dysarthric intelligibility that emulates human listening tests.

Chapter 6 presents a novel pronunciation assessment method that models healthy pronunciation with HMMs to detect pronunciation errors in speakers with dysarthria.

Finally, Chapter 7 concludes the thesis and discusses future work.

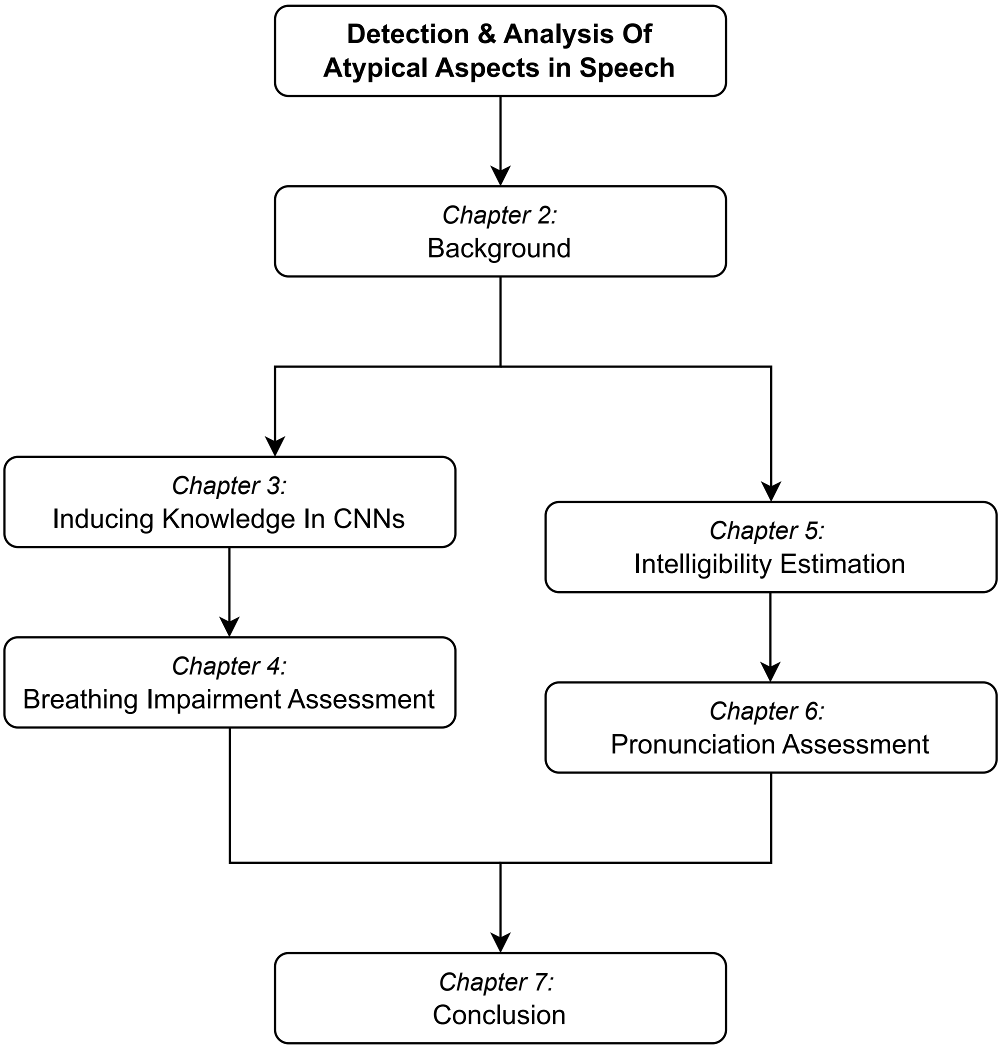

Detection & Analysis Of Atypical Aspects in Speech
Chapter 2: Background
Chapter 3: Inducing Knowledge In CNNs
Chapter 4: Breathing Impairment Assessment
Chapter 5: Intelligibility Estimation
Chapter 6: Pronunciation Assessment
Chapter 7: Conclusion

Figure 1.2: Schematic overview of this thesis.

# 2 Background

This chapter introduces topics, that provide a background for the following chapters. We start by explaining some re-occurring terms (cf. Section 2.1), then give a short overview of how speech is produced in Section 2.2 and conclude with Section 2.4 on introducing some popular approaches to atypical/paralinguistic speech assessment.

## 2.1 Terminology

In the upcoming chapters, some terms will be used that we would like to introduce beforehand:

**Atypical speech:** Speech that deviates from normal healthy speech. Atypical aspects are noticeable aspects that go beyond the large degree of variability of speaking styles (Stemmer et al., 2010).

**Paralinguistics:** A part of meta-communication that refers to the non-verbal aspects of communication, such as prosody, emotions, and other temporary speaker states and traits. Paralinguistics lies at the intersection of speech and social sciences.

**Intelligibility:** Can be defined as what is understood by the listeners of a phonetic realization (Yorkston et al., 1996). In the context of disorders, De Bodt et al. (2002) defines it as a combination of the main dimensions of speech production: voice quality, articulation, nasality, and prosody. Intelligibility is also used in the context of speech transmission or speech synthesis.

**Phonation:** The process of passing air from the lungs through the vocal folds to produce speech sounds through quasi-periodic vibration (Titze and Martin, 1998).

**Phoneme:** A unit of sound of a language that allows distinguishing one word from another. The phonemes of a language constitute the minimal set of symbols needed to describe the pronunciations of all words in that language.

**Phone:** An acoustic realization of a speech sound. A Phone is any distinct speech sound and, as opposed to phonemes, not specific to any language.

**Phonetic transcription:** This form of transcription uses a sequence of phones to describe the uttered speech.

**Articulatory features:** Articulatory features are characteristics of phones that refer to how the articulators are positioned during the production of a phone (cf. Section 2.2). They can be measured by means of an articulograph, but are in this thesis only derived from acoustic-articulatory prediction.

## 2.2 Speech production – phonetic summary

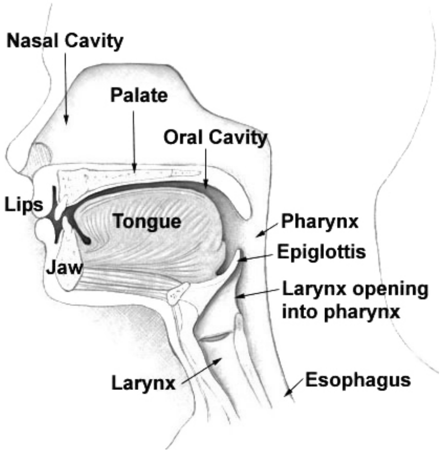


Figure 2.1: Human vocal apparatus used to produce speech (Wikimedia, 2023).

Figure 2.1 illustrates the human vocal apparatus used to produce speech. Speech production is a complex process that involves many body parts, yet is mostly controlled subconsciously. Humans self-assess their speech through hearing; together, production and perception create a so-called feedback loop. For accurate articulation, many body parts have to cooperate correctly to transfer a message. Air from the lungs is sent, creating the right amount of air pressure, to create the right air vibration (called phonation) through the trachea to the larynx. The larynx has two vocal folds (Story, 2002). During respiration, the vocal folds are in an abducted position (separated), leaving a gap between them which is called the glottis. During phonation, the vocal folds adduct (move together), narrowing the glottis. This makes the subglottal pressure build up below the vocal folds. When the pressure is high enough, the vocal folds are forced to separate and the air stream is allowed to flow through the process. Once the pressure has dropped, the vocal folds close back together. The subglottal pressure then builds up again. This repeated process of opening and closing the glottis is called the glottal cycle and is the basis of phonation. The rate of closure is typically referred to as fundamental frequency. The fundamental frequency (F0) of the voice can be altered by the muscles surrounding the vocal folds. However, when the vocal folds are in an abducted position, they do not vibrate and there is no phonation, and unvoiced sounds can be produced (Titze and Martin, 1998).

The wave that passed through the vocal folds next goes through the vocal tract. It passes through the laryngeal cavity, the pharynx, the oral cavity, and the nasal cavity. The vocal tract modulation shapes the sound wave and determines the resonances of the sounds, called formant frequencies. The most conscious modulation is happening in the oral cavity, with the jaw, tongue, and lips.

**Articulation:** The outcoming sound depends on the shape of the cavities and on how much and where they are narrowed. These are broadly categorized into manner of articulation, degree or height of constriction, and the place of constriction determines the place of articulation (Ladefoged and Johnson, 2014). Naturally, articulation is continuous and not discrete but it helps us to describe the sounds. An overview can also be found in the Appendix Table A.1. In the following, we will give a short overview of the most important articulatory categories, phonemes are denoted in ARPABET notation (Klautau, 2001).

**Manner of articulation:** describes the way of constriction in the pharyngeal and oral cavity when producing consonants. Several categories can be distinguished (Ladefoged and Johnson, 2014): *Plosives* are stop sounds; the airflow is shortly interrupted, followed by a short burst of air release, e.g. /p/, /b/, /k/, /t/, /d/. *Fricatives* are sounds characterized by a turbulent air stream caused by a small gap in the vocal folds. Examples are /f/, /v/, /s/, /z/, /S/, /Z/. *Laterals* are sounds by which an air stream passes along one or both sides of the tongue. An example is /l/. *Trills* are sounds characterized by vibrations between the tongue and the place of articulation, e.g. /r/. *Nasals* main characteristic is a lowered velum that connects the oral and nasal cavities creating resonances. Examples of nasals are /m/, and /n/. *Approximants* are produced with articulators approaching each other to create turbulent airflow, and closely resemble vowels. Examples are /j/ or /w/.

**Place of articulation:** or the place of constriction in the vocal tract, can be distinguished into the following categories: *labial*, where the lips make contact, e.g. /m/, /p/ or /b/. *labiodental*, a constriction between the lower lip and teeth, e.g. /f/ or /v/, *dental*, where the tongue touches the teeth. e.g. /th/, *alveolar*, where tongue tip touches the alveolar ridge, e.g. /t/, /s/, /n/, /d/, /l/, /z/, /r/, *retroflex* are sounds, where the tongue has a curled shape backward behind the alveolar ridge, e.g. /r/ and *velar* are sounds, where the back part of the tongue touches the velum, e.g. /N/,/k/,/g/, /r/, /x/ and /G/.

**Height of articulation:** When producing vowels, the resonance frequencies of voiced sounds are modulated by the vocal tract, oral and nasal cavity. Oral articulators are place of an elevation of the tongue, the rounding of the lips, and the opening of the velum. Vowels with a high position of the tongue and narrow opening between the pharynx and oral cavity include /i/ or /y/. Examples of vowels with a low position of the tongue and wide opening are /a/ or /o/. For vowels, four heights are commonly distinguished: high, mid-high, mid-low, and low.

## 2.3 Atypical aspects in speech

This thesis covers studies on speech which differs from normal healthy normal, hence atypical speech. These atypical aspects can stem from a variety of causes, such as neurological diseases, and psychological disorders, such as depression. In the following, we will give a high-level overview of the types of atypical speech in this thesis:

**Parkinson's disease** is a neurodegenerative disease that affects the motor system. It is characterized by tremors, rigidity, bradykinesia, and postural instability (Bloem et al., 2021). The speech of patients with Parkinson's disease is often characterized by a monotone voice, reduced loudness, reduced pitch range, reduced speech rate, reduced prosody, reduced intelligibility, and reduced articulation. To date, there is no cure, but there are treatment options, such as medication, which could be monitored by speech analysis. Overall severity is commonly evaluated on the Unified Parkinson's Disease Rating Scale (UPDRS) (Goetz et al., 2008) which only contains one item on speaking capabilities. Speaking deficits are commonly rated on the modified version of the Frenchay Dysarthria Assessment (m-FDA) scale (Vásquez-Correa et al., 2018), in which a total of 13 items are evaluated. Seven aspects of the speech are rated, including breathing, lips movement, palate/velum movement, laryngeal movement, intelligibility, and monotonicity.

**Alzheimer's disease** is a neurodegenerative disease that affects the memory and cognitive functions (Scheltens et al., 2016). Alzheimer's disease is known to rather affect language, resulting in reduced vocabulary, more hesitations reduced speech rate. To date, there is no cure, but medication to alleviate the symptoms could be monitored by speech analysis. AD is frequently assessed on the Mini-Mental State Examination (MMSE), a 30-point questionnaire on daily tasks to measure cognitive impairment (Folstein et al., 1975).

**Sleepiness** is a temporary speaker state, which can occur in healthy and sick people. Although sleepiness is a multi-modal phenomenon, it manifests in speech. Typically, sleepiness affects articulation and leads to slurred, less crisp pronunciation, mispronunciation, and effects on speech quality such as tensed, nasal, or breathy speech.

**Lip filler** surgery is a cosmetic procedure that is used to increase the volume of the lips. The procedure is performed by injecting a filler, such as hyaluronic acid into the lips. After surgery, speaking needs to be adjusted and could lead to less crisp plosives. Medical professionals are interested in the success of the surgery, which can only be performed after local anesthesia has worn off.

**Dysarthria** is a collective term for neurological motor speech disorders. It is broadly characterized by reduced articulation, reduced intelligibility, reduced speech rate, and reduced prosody. Dysarthria can be caused by a variety of neurological diseases, such as Parkinson's disease, stroke, multiple sclerosis, cerebral palsy, muscular dystrophy, brain injury, brain tumor, and others. Speech intelligibility assessment of people with dysarthria, typically performed by therapists, could be supported by automatic methods (Duffy, 2012).

**Voice disorders** are a collective term for disorders of the voice (Ramig and Verdolini, 1998). They are broadly characterized by reduced loudness, reduced pitch range, reduced prosody, reduced intelligibility, and reduced articulation. Voice disorders can be caused by a variety of neurological diseases, such as Parkinson's disease, stroke, multiple sclerosis, cerebral palsy, muscular dystrophy, brain injury, brain tumor, and others. Speech intelligibility assessment of people with voice disorders, typically performed by therapists, could be supported by automatic methods.

**Cleft lip and palate** is a congenital opening in the upper lip and palate. It can be treated with surgery, but it requires speech therapy, as people's speech is often characterized as breathy, hoarse, nasal, and low intensity (Schuster et al., 2006).

**Laryngectomy** is a surgical procedure to remove (parts of) the larynx due to cancer. As a result, subjects can have trouble breathing, swallowing, and producing voiced sounds which has a significant impact on intelligibility. However, the larynx can be partially restored. With the help of speech therapy, rehabilitation options include esophageal speech, tracheoesophageal speech, and the use of an electrolarynx (Pereira da Silva et al., 2015).

## 2.4 Literature overview of atypical speech ananlysis

In atypical speech analysis, the relevant information is often overlaid on the content of the message. Besides prior knowledge, it is often not clear, what information is relevant for the task or may differ across classes/labels. Figure 2.2 illustrates the standard building blocks of speech analysis. Approaches can be differentiated by the speech material used, e.g. spontaneous speech, sustained vowels, or read text. Regarding feature extraction, we broadly differentiate conventional speech analysis (cf. Section 2.4.1) and neural network-based approaches (cf. Section 2.4.2). For an assessment, features are input to a classifier or regressor. Choices range from simple methods, such as k-nearest neighbors (Arjmandi and Pooyan, 2012), or Gaussian mixture models (GMMs) (Dibazar et al., 2002; Godino-Llorente et al., 2017), The most popular choice are support vector machines (SVMs) (Arjmandi and Pooyan, 2012; Bocklet et al., 2013), since they perform well at low sample sizes and high-dimensional vectors or random forests (Noroozi et al., 2017). Neural networks have also been used as classifiers, e.g. by (Berus et al., 2018).

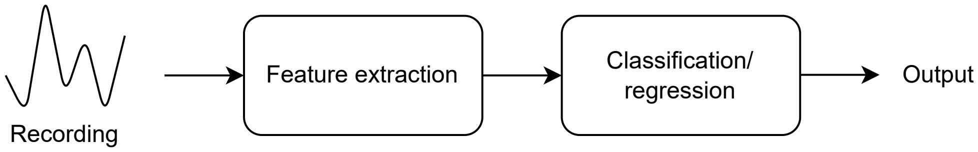


Figure 2.2: Simplified schematic of speech analysis.

### 2.4.1 Conventional speech analysis

One challenge was and is to find a fitting representation for a task. Most conventional speech analysis methods are based on spectrum short-term spectrum-based processing. A crude solution has been acoustic low-level descriptors, such as mel-frequency cepstral coefficients (MFCCs) or perceptual linear prediction (PLP) coefficients (Hönig et al., 2005). Tanner et al. (2005) use statistics of long-term average spectra to assess voice characteristics before and after dysphonia treatment, which was extended by Smith and Goberman (2014) to measure voice characteristics of Parkinson's patients. To assess the changes in speech production from Parkinson's disease, Skodda et al. (2011) calculated the values of the first two formants from each vowel to analyze the vowel articulation by computing triangular Vowel Space Area (tVSA). tVSA has also been used to detect vowel articulation problems after oral surgery in (van Son et al., 2018). In (Vásquez-Correa et al., 2017b), Parkinson's disease was classified by analyzing the relationships between tVSA and other acoustic features from sustained vowel articulation. Relevant information can also lie at the transition of sounds: In (Orozco-Arroyave et al., 2015), the authors analyze voiced-unvoiced transitions to detect Parkinson's disease in spontaneous speech, the segments are modeled with MFCCs and spectral energies and fed into an SVM.

Since most short-term processing methods yield features for short segments, they need to be aggregated into a per-utterance representation. Additionally, data is typically labeled on a per-utterance basis or per-speaker basis. Commonly, statistical functionals such as mean, standard deviation, skewness, kurtosis, and percentiles are computed, e.g. in (Bocklet et al., 2013; Orozco-Arroyave et al., 2014b). A very exhaustive solution was proposed by Schuller et al. (2013) that is typically referred to as the ComParE 2013 features: A feature set composed of 4 energy-related features (e.g. zero-crossing-rate), 55 spectral features (e.g. energy of different frequency bands), 6 and voicing features (e.g. F0, jitter and shimmer) can be aggregated into a per utterance representation by calculating a large variety of statistical functionals, forming a 6373-dimensional vector. This representation still serves as a baseline for paralinguistic tasks such as emotion recognition (Trigeorgis et al., 2016), Parkinson's disease (An et al., 2015) and even COVID-19 detection (Schuller et al., 2021). This idea was refined into the Geneva Minimalistic Acoustic Parameter (GeMAPS) feature set. An extended version, eGeMAPS, that contains cepstral features and formant-based features totals 88 features (Eyben et al., 2016). As per the name, this more minimalistic feature set contains 18 features, of which statistical functionals are computed, resulting in a 62-dimensional vector. Besides statistical functionals, utterance-level representations can be derived from bag-of-audio-word (BoAW) (Liu et al., 2010; Schmitt and Schuller, 2017). BoAW representations are typically derived from a clustering algorithm, such as k-means, with a fixed number of centroids. This is followed by vector quantization: Features are built by creating histograms of feature vectors' closest centroids. BoAW representations are robust, time-invariant, and non-reconstructable, which is good for privacy. i-vectors are another popular representation. Originally developed for speaker verification, i-vectors follow a generative approach: MFCCs are modeled with a Gaussian mixture model (GMM) i-vectors are then extracted through total variability analysis (Dehak et al., 2010). They were used for sleepiness estimation in (Ravi et al., 2019).

### 2.4.2 Neural network based speech analysis

More recently, deep learning-based methods have been applied for **representation learning**: x-vectors are utterance-level speaker embeddings from a neural network that is trained on speaker classification (Snyder et al., 2018). Another popular approach is auto-encoder embeddings: In (Freitag et al., 2017), a sequence-to-sequence denoising RNN-autoencoder called auDeep-based embeddings is proposed. The authors propose to average frame-level representations to obtain an utterance-level representation. In (Vasquez-Correa et al., 2020), a CNN-based auto-encoder is trained on healthy speech data. These auto-encoder embedding representations are then extracted for training a separate pathological speech classifier. Another category of representations is pre-trained self-supervised embeddings, that were developed for speech recognition, such as wav2vec 2.0 (Baevski et al., 2020). Similarly, frame-level embeddings have to be aggregated into utterance-level representations before they are fed to a classifier.

For **end-to-end modeling**, one main challenge is to avoid over-fitting, since neural networks are 'data-hungry'. Common neural-network regularization is done with dropout, pooling, or residual connections. In Bhati et al. (2019), LSTM Siamese networks were proposed for pathological speech detection. Siamese networks are trained on pairs of input and share weights in the first layers, which is beneficial to extract features that are problem-discriminative and robust to non-relevant information. Another option is reducing the number of trainable parameters. Since CNNs are common in speech analysis, dilated convolutions can be used instead (Yu and Koltun, 2015): As applied in Li et al. (2019) for emotion recognition, dilated convolutions use the size receptive field, but skipping values. A recent line of research is fine-tuning pre-trained models. This was proposed for example for emotion recognition in Chen and Rudnicky (2021).

## 2.5 Summary

In this chapter, we overviewed areas of research that are relevant to our work. We began with re-occurring terminology in Section 2.1, followed by a description of the speech production mechanism and articulation in Section 2.2. In Section 2.3, we gave the reader a basic idea of the different types of atypical speech that are part of this thesis. In Section 2.4, we overviewed the literature on atypical speech analysis.

# 3 Inducing knowledge into raw waveform CNNs

Neural network-based atypical speech assessment can be carried out with different architectures. System choices, such as type of input, layer type, number of layers, and cost function can lead to a variety of model characteristics. In this chapter, we opted for a neural network framework that assumes minimal prior knowledge about the problem. By directly using the raw waveform as input, the CNN network can learn to extract features instead of feeding hand-crafted features, e.g. spectrograms, which may not perfectly suit the task. This approach has also been widely adopted for representation-learning, e.g. for wav2vec (Schneider et al., 2019).

In this chapter, we investigate the impact of different design choices. To effectively estimate speech production parameters, different aspects of the raw waveform CNN framework can be exploited. We propose different methods, that improve efficiency by following different intuitions about a certain speech production parameter to be estimated.

In Section 3.2 we investigate integrating prior knowledge to improve Alzheimer's assessment. In Section 3.3, we investigate integrating phonetic knowledge as a pre-training scheme for sleepiness estimation. The networks trained in this section will also be used in Chapter 5.

## 3.1 Raw waveform convolutional neural networks

In this section, we describe the raw waveform CNN architecture used in the following chapters. Originally developed for speech recognition (Palaz et al., 2013), raw waveform 1D-CNN is a versatile architecture. A similar architecture has been applied to other tasks, such as speaker verification (Muckenhirn et al., 2018), speaker recognition (Ravanelli and Bengio, 2018), gender identification (Kabil et al., 2018), detection of spoofing of speaker verification systems (Muckenhirn et al., 2017) or depression detection (Dubagunta et al., 2019). Additionally, popular raw waveform CNN applications include wavenet (van den Oord et al., 2016), and recently, it is applied in self-supervised learning for models such as wav2vec (Schneider et al., 2019). In the following, we will demonstrate, how the raw waveform CNN architecture can be used for an atypical speech assessment.

In Figure 3.1 we illustrate the framework: the network consists of a filter stage, which itself consists of $n$ convolution layers (conv), maximum pooling (maxp) and relu activations followed by a multilayer perceptron (mlp). At the output, the CNN predicts a score per input segment, where for (i) **classification**, a sigmoid output is used for binary classification and a softmax output layer for multi-class problems, which results in a posterior probability, for which we typically compute the loss with a (binary) cross-entropy loss function, and for (ii) **regression**, a linear output layer is used typically, which results in a score, computing the loss from a mean squared error (mse) loss function. We train randomly initialized models; a batch-size of 256 has proven a robust parameter. We mainly used the Adam optimizer (Kingma and Ba, 2014). For training, we use a decaying learning schedule which halves the learning rate between $10^{-3}$ and $10^{-7}$ whenever the validation loss stopped reducing. For every input segment, the model outputs a score. These scores are then averaged to get the per-utterance score.

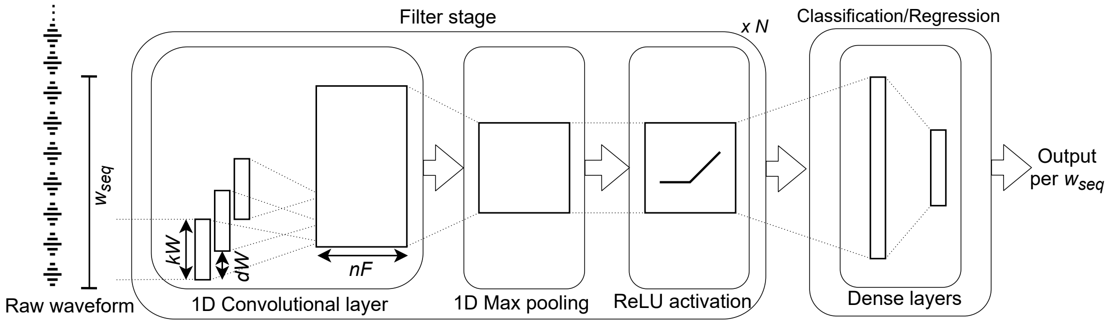


Figure 3.1: Illustration of the proposed CNN architecture. $w_{seq}$: input segment, $N_f$: number of filters, $kW$: kernel width, $dW$: kernel shifts.

Figure 3.1 also illustrates the processing at the first convolution layer. $kW$ denotes the kernel width in samples, $dW$ denotes the stride or kernel shift in samples, $w_{seq}$ is the segment of speech that is processed at one time frame and $nF$ is the number of filters in the convolution layer. In Palaz et al. (2019); Muckenhirn et al. (2018), it has been found that, by modifying $kW$, different information related to the speech production mechanism can be learned. More precisely, if $kW$ covers a signal length of about 20 ms (segmental), the first convolution layer tends to model voice-source-related information. Similarly, if $kW$ covers a signal of about 2 ms of length (sub-segmental), the first convolution layer tends to model vocal tract system-related information, such as formant information. We chose the input segment length as $w_{seq}$ = 250ms (the average length of a syllable) and the shift as $dw$ = 10ms unless specified otherwise.

**Architectures:** Table 3.1 presents the architectures used. We differentiate based on the first convolution layer kernel width: Depending upon the length of the filters in the first convolutional layer, we distinguish (a) sub-segmental modeling (subseg), where $kW$ = 30, span over $2ms$, equivalent to less than 1 pitch period, and (b) segmental modelling (seg), where

$kW = 300$ spanning $20ms$, equivalent to 1 to 5 pitch periods. The AF-CNN architecture uses sub-segmental modeling, see Section 3.3. The classification stage consists of one hidden layer with 100 units.

Table 3.1: CNN architectures. $N_f$: number of filters, $kW$: kernel width, $dW$: kernel shifts, $MP$: max-pooling.

| **Model** | Layer | Conv | | | MP |
|---|---|---|---|---|---|
| | | $N_f$ | $kW$ | $dW$ | |
| subseg | 1 | 128 | 30 | 10 | 2 |
| | 2 | 256 | 10 | 5 | 3 |
| | 3 | 512 | 4 | 2 | - |
| | 4 | 512 | 3 | 1 | - |
| seg | 1 | 128 | 300 | 100 | 2 |
| | 2 | 256 | 5 | 2 | - |
| | 3 | 512 | 4 | 2 | - |
| | 4 | 512 | 3 | 1 | - |

## 3.2 Alzheimer's assessment

Alzheimer's disease primarily affects cognitive functions (cf. 2.3), and has successfully been assessed through analysis of language, typically manual transcripts (cf. Section 3.2.1). In this section, we investigate the potential of raw waveform CNNs to assess Alzheimer's disease from speech and whether text-based systems can benefit from a fusion.

Dementia is a neurodegenerative disease and cause of major disability in the elderly population worldwide, with at least 10 million new cases reported every year (World Health Organization, 2018). Alzheimer's Disease (AD) is the most common cause of dementia (World Health Organization, 2018; Alzheimer's Association, 2017). Automatic early diagnosis systems promise to help alleviate this societal burden with timely and optimal management.

In the following, some experiments were conducted as part of a collaboration with project partners from the ***T**raining Network on **A**utomatic **P**rocessing of **PA**thological **S**peech* (TAPAS) – a Horizon 2020 Marie Skłodowska-Curie Actions Innovative Training Network European Training Network. In (Cummins et al., 2020), we compared different contemporary acoustic- and linguistics-based systems and explore combining the information learned and the potential gains of multi-modal systems.

### 3.2.1 Related works

The literature on Alzheimer's classification can be broadly divided by modality; text-based approaches have been proven successful. In an early work by Fraser et al. (2016) the authors show that linguistic features yield stronger performances in Alzheimer's detection. In (Fritsch et al., 2019), we showed that evaluating transcripts on two different language models, one built from control and one from Alzheimer's transcripts, yields high discriminatory power. Other linguistic systems utilize *Global Vectors* (GLoVe) word embeddings (Pennington et al., 2014) or hierarchical attention neural network (Yang et al., 2016).

Among the submissions to the 2020 Interspeech ADReSS challenge, the best performing system (Yuan et al., 2020) proposed using BERT and ERNIE models that were fine-tuned to perform the classification task. In addition to the sequence of word embeddings, a forced alignment was used to add symbols for different pause lengths, which at a granularity between 3 and 6 different symbols improved performance (e.g. short, medium, and long pauses). In (Koo et al., 2020) propose a comparison of pre-trained word embeddings, of which Transformer-XL yields better performance than the openAI's GPT, even though being significantly bigger. Overall, word-level embeddings from pre-trained language models have been used by a majority of participants, while not much creativity was shown in using the audio data.

For acoustic systems, in paralinguistic scenarios, baselines frequently include *Bag-of-Audio Words* (BoAW) system (Schmitt and Schuller, 2017). Additionally, we propose to test a *Siamese network* (Bromley et al., 1994) as well as *Convolutional Neural Network*. At the time of publication, these three systems have not been used for Alzheimer's recognition.

### 3.2.2 Proposed approach

We propose to compare three different state-of-the-art acoustic methods as well as two text-based methods. Additionally, we examine if a fusion between acoustic and text-based systems yields performance improvements, which would mean that they model complementary information.

Among the three acoustic approaches, we proposed were: i) MFCC-based BOAW, that have proven to robustly summarize acoustic characteristics on an utterance level, ii) SiameseNet, which share CNN layers for feature extraction and that is trained with a contrastive loss, iii) a 1-D CNN end-to-end system that is trained on raw waveform, as described in Section 3.1. On top of feeding the 1-D CNNs the raw signals, we propose to guide the network toward focusing on voice source information. Towards that, we filter the signal with a zero frequency filter (cf. Section 3.2.3). This filtering technique has already been successfully applied to depression detection in Dubagunta et al. (2019).

We deployed two text-based systems: (i) a bidirectional LSTM network followed by an attention layer (bi-LSTM-ATT) and (ii) a bi-directional hierarchical attention neural network (bi-HANN), which is motivated by the hierarchical structure of language applies attention layers at word and at sentence layers is a state-of-the-art text-based method (Pan et al., 2019).

### 3.2.3 Zero frequency filtering

Zero frequency filtering (ZFF) is an algorithm designed for epoch extraction (Murty and Yegnanarayana, 2008). An epoch is the instant of significant excitation of the vocal folds, meaning in the glottal closure instance (in voiced sounds). Theoretically, the impulse-like excitation characteristics in the time domain spread across all the frequencies equally well represented in the frequency domain. The occurrence of the excitation impulses can therefore be found as a deviation from the center frequency in a narrowband. The advantage of choosing a zero frequency resonator filter is that the time-varying vocal-tract system does not affect the discontinuities at 0 Hertz. The algorithm aims to remove these time-varying characteristics of the vocal tract information from the signal that way, instead of modeling this information explicitly. By putting two cascaded resonators at 0 Hz, the effect of vocal tract resonances is minimized and the excitation signal can be extracted. A trend removal operation is necessary to counteract the effects of the exponential growth/decay of the two 0 Hz resonators.

### 3.2.4 Database

The ADReSS challenge organizers provided audio data and corresponding transcripts performing the Cookie Theft picture description task as well as meta-data such as age and gender (Luz et al., 2020). The organizers provided a train and test set. Unfortunately, no standard development set was proposed, which is why participants that proposed less compute-heavy systems reported leave-one-speaker-out cross-validation results of the training set, whereas others had to create their own folds for cross-validation. The challenge proposes a binary classification task of AD, evaluated in terms of accuracy and F1-score, and a regression task of estimating the patients' score on the Mini-Mental State Exam (MMSE) (Folstein et al., 1975), evaluated in terms of RMSE. The audio data was available in two different forms: full speech files and segmented speech chunks. The segmented chunks were generated by applying a log-energy threshold-based voice activity detector. The average length of a full recording is 2min24 and that was segmented into on average 24.86 segments.

Table 3.2: ADReSS database statistics.

| | **Train** | **Test** |
|---|---|---|
| Gender [m/f] | 24/30 | 11/13 |
| Age | 66.9±6.3 | 66.4±6.6 |
| MMSE | 17.0-±5.5 | 19.5±5.3 |
| Length | 2h09 | 1h06 |

### 3.2.5 Systems

We compare our proposed approaches to the best baselines from the challenge organizers. The best performing acoustic system was ComParE features (cf. 2.4) fed into an SVM/SVR with rbf kernel for classification/regression respectively. The best-performing linguistic system is a feature set based on parts-of-speech tags that are again fed into an SVM/SVR with rbf kernel.

In our comparison of acoustic systems, we propose to use three different approaches: (i) MFCC-based Bag-of-Audio-Words (BoAW) with 125 centroids, that are fed into an SVM/SVR with a linear kernel for classification/regression task respectively, (ii) a CNN-based Siamese Network trained on 16-second input segments of log-Mel spectrograms and (iii) raw waveform CNNs with 250ms input segments of either raw audio or ZFF signals. We train our models on the full recordings. The architectures used can be found in Table 3.1. As described in Section 3.1, we use a sigmoid output and a binary cross-entropy loss function for the classification task and linear output with MSE loss.

The two text-based approaches are (i) bi-LSTM-ATT and (ii) bi-HANN. Both used 100-dimensional GloVe word embeddings and a maximum word number of 200 per transcript to obtain a fixed-length representation.

### 3.2.6 Results

In Table 3.3, we compare our systems to the challenge baseline systems on a 10-fold cross-validation. On the classification task, most of our proposed acoustic systems outperform the baseline both on train and test set, while the acoustic baseline has a strong RMSE result of 6.14. We observe that the SiameseNet performs similarly to our own proposed raw waveform CNNs. Raw waveform CNNs compare well to the other acoustic methods. Our main finding was that ZFF signals showed a better performance than raw audio, even though Alzheimer's is not known to affect voice source, but has effects on prosody, e.g. in terms of self-consciousness and emotions. As expected, the text-based bi-LSTM-Att and bi-HANN networks outperform all acoustic methods. In the last row, we present the result of a late fusion. For the classification task, we propose a majority voting of our four best systems. For regression, we proposed a weighted average of scores, where the weights are determined on the development set. On the test set, for both classification and regression, late fusion yielded performance gains over the linguistic system, especially for the classification result.

Overall, in our comparison, acoustic get outperformed by linguistic methods, which is unsurprising given human transcripts are provided. Small gains were found when fusing acoustics and linguistics approaches.

### 3.2.7 Increasing input segment length

To explore the full potential of raw waveforms for Alzheimer's recognition, this work was continued to improve performance. In the previous section, we used a sub-segmental architecture with input segments of 250ms of ZFF signals. Based on this result, we explore longer input segments. As shown in Table 3.4, it was found that at the same configuration, an input segment length of 4 seconds yields optimal performance for the classification task. Presumably, longer input lengths better capture prosodic differences in Alzheimer's language. These results were

Table 3.3: A comparison of the proposed approaches on the ADReSS Challenge training and test set. Results are the average performance across a nine-fold cross-validation.

| | **Dev** | | | **Test** | | |
|---|---|---|---|---|---|---|
| **Baselines** | **Acc.** | **F1** | **RMSE** | **Acc.** | **F1** | **RMSE** |
| ComParE | .565 | - | 7.29 | .625 | - | 6.14 |
| Linguistic | .768 | - | 4.38 | .750 | - | 5.20 |
| **Approach** | | | | | | |
| $\text{BOAW}_{125}$ MFCC | .630 | .623 | 7.05 | .563 | .561 | 6.88 |
| SiameseNet | .628 | .731 | – | .708 | .708 | – |
| *Raw signal CNN subseg* | .652 | .721 | 10.1 | .657 | **.731** | 12.05 |
| *Raw signal CNN seg* | .713 | .762 | 8.89 | .562 | .667 | 8.93 |
| *ZFF signal CNN subseg* | **.741** | **.780** | 7.58 | **.667** | .692 | **6.67** |
| *ZFF signal CNN seg* | .684 | .751 | **7.58** | .583 | .643 | 6.75 |
| bi-LSTM-Att | **.842** | **.842** | 5.49 | .813 | .812 | 4.66 |
| bi-HANN | .827 | .826 | **4.86** | .729 | .726 | 4.74 |
| Fusion - Maj. / W-avg | .831 | .829 | 7.64 | **.852** | **.854** | **4.65** |

published in another fusion study (Villatoro-Tello et al., 2021).

Table 3.4: ADReSS test performance with a sub-segmental architecture and ZFF signal input at different input segment lengths.

| | **Dev** | | | **Test** | | |
|---|---|---|---|---|---|---|
| Input length | Acc | F1 | RMSE | Acc | F1 | RMSE |
| .250s | .741 | .780 | 7.58 | .667 | .692 | **6.67** |
| 1s | .703 | .754 | 7.91 | .667 | .652 | 9.12 |
| **4s** | .722 | .764 | 7.67 | **.792** | **.821** | 7.06 |
| 8s | .750 | .787 | 8.73 | .625 | .667 | 8.26 |

In (Villatoro-Tello et al., 2021), we fused acoustic systems with a new text-based approach: subjects' lexical availability of a class is quantified on a per-transcript level. Lexical Availability (LA) is psycho-linguistics-inspired and identifies the most accessible vocabulary of the interviewed subject. A feature vector is constructed from the available lexicon of subject, as well as its correlation to the average representation for a class. Notably, this representation is not neural network-based. When feeding these representations to a logistic regressor, an accuracy of .87 was achieved. When late-fusing of these results and the best result reported in Table 3.4, an improved accuracy of .90 could be achieved, which is highly competitive with the results reported by other challenge participants. Within this publication, in addition to Alzheimer's this assessment, the same experimental setup was applied to a depression detection task and yielded a significant improvement. This validates the approach's usefulness on cognitive tests.

## 3.3 Integrating speech production knowledge

Speech-based degree of sleepiness estimation is an emerging research problem. In this section, we investigate raw waveform CNNs to estimate the degree of sleepiness. Within this approach, we also propose the integration of speech production knowledge through transfer learning.

### 3.3.1 Related works

Assessing sleepiness is relevant in scenarios, such as in preventing accidents or in evaluating when to recommend a break. Furthermore, sleep deprivation increases the mortality risk. To put this relevance into perspective: in 2016, the American think tank RAND reported an estimated US$138 billion damage to the Japanese economy (2.92% of its GDP) caused by sleepiness at work, which is why companies, among other things, offer incentives to sleep more than six hours per night (Hafner et al., 2016). Although sleepiness is a multi-modal phenomenon, speech is one of the cheapest modalities that can be captured, most notably in a non-intrusive manner. Sleepiness can be subjectively assessed on the Karolinska Sleepiness Scale (KSS) (Shahid et al., 2011), which ranges from 1 (extremely alert) to 9 (very sleepy) in steps of one. In this work, we focus on developing objective or automatic methods to predict sleepiness.

In the literature, estimating sleepiness has been addressed by investigating acoustic factors. Traditionally, baseline systems used a large number of general-purpose low-level descriptors (LLDs) such as short-term energy, short-term spectrum, voice-related features, and their functionals, as in (Schuller et al., 2019). In (Schuller et al., 2014), Schuller et al. reviewed contributions to the Interspeech 2011 Speaker State Challenge on sleepiness estimation, which is labeled in terms of KSS. Sleepiness is considered a medium-term speaker state, meaning effects that usually last a few hours. It is expected to generally affect motor coordination processes and the cognitive processing of speech. This manifests in terms of changes in prosody such as monotonic and flattened intonation, in shifted speech rate (Krajewski et al., 2010; Vogel et al., 2010), in articulation, such as slurred, less crisp pronunciation, mispronunciation (Bratzke et al., 2007) and in speech quality such as tensed, nasal, or breathy speech (Kostyk and Rochet, 1998). Hönig et al. (2014) analyzed the LLDs extracted from the Interspeech 2011 Speaker State Challenge sleepiness data. They found that male sleepiness correlated more with spectral changes such as less canonical pronunciation, whilst female sleepiness correlated more with lowered $F_0$.

More recently, as part of the Interspeech 2019 ComParE challenge, histogram representations of clustered LLDs, known as bag-of-audio-words (BoAW), and feature representations from sequence-to-sequence auto-encoders (S2SAE) trained on Mel-spectrograms were studied (Schuller et al., 2019). In (Gosztolya, 2019) the authors created utterance-level Fisher vectors by training a GMM on frame-level MFCCs, which are used for classification with SVM. Similarly, Wu et al. (2019a) investigated the extraction of Fisher vectors from a large variety of acoustic features. In (Elsner et al., 2019), the authors investigated raw waveform CNNs including data augmentation, such as inputting reverse samples, adding noise or using noisy labels. Yeh et al. (2019) presented a system that uses frame-level eGeMaps features that were input to a BLSTM-CNN network with attention. For data augmentation, an adversarial auto-encoder was used to generate synthetic samples.

Additionally, border cases, e.g. samples with low and high KSS scores, that are intuitively more relevant to detect, were selected for an additional classifier to be used for score fusion. Wu et al. (2019b) aimed to address the ordinality of the KSS labels and introduce an ordinal triplet loss that is used to train binary classifiers for each label individually. Ravi et al. (2019) used between-frame entropy, a measure that correlates with speech rate, to detect outliers, and create utterance-level iVectors from voice quality features.

While the above-mentioned contributions investigated many relevant acoustic aspects and address issues such as the ordinality of the KSS labels or the imbalance of a data set to solve sleepiness estimation as a classification/regression problem, acoustic-phonetic changes in sleepy speech have not yet been considered. Therefore, our goal is to study whether speech production differences from a phonetic perspective can be captured for degree of sleepiness estimation, inspired by Dubagunta and Magimai-Doss (2019).

### 3.3.2 Proposed approach

We propose to estimate sleepiness with raw waveform CNNs. Our initial experiments are performed with the raw waveform framework (cf. Section 3.1). To tackle data scarcity, we develop a pre-training method that induces phonetic knowledge into CNNs. Intuitively, a model that is trained is able to predict articulatory features from speech should perform well on sleepiness estimation. Based on common phonetic representation (cf. Section 2.2), we pre-trained four models, namely on (i) height of articulation, which refers to the position of the tongue, (ii) manner of articulation refers to the way of constriction in the pharynx and oral cavity, (iii) place of articulation refers to where the constriction is happening, and (iv) vowels, is a model that only predicts vowels. We propose to fine-tune these models as well as combining their output scores to find out if they model complementary information.

### 3.3.3 Data and experimental protocol

The continuous sleepiness sub-challenge corpus was part of the Interspeech 2019 ComParE challenge (Schuller et al., 2019). The corpus is also referred to as the Düsseldorf Sleepy Language Corpus and was created at the Institute of Psychophysiology, Düsseldorf and the Institute of Safety Technology, University of Wuppertal, Germany. It consists of 5564 utterances (5 hours 59 minutes) in the training set, 5328 utterances (5 hours 44 minutes) in the development set, and 5570 utterances (5hours 58minutes) in the test set from a total of 915 subjects (364 females, 551 males) aged 12 to 84 years with a mean age of 27.6 ± 11.0. Recordings happened between 6 pm and midnight; each speaker provided between 15 minutes and

1 hour of speech. No speaker IDs, speaker gender, or age information are provided. Speech data consists of different reading and speaking tasks as well as narrative speech. The material was recorded in quiet rooms using a microphone/headset/hardware setup and providing a prompt with the tasks on a computer. According to the KSS scale, the labels range from 1 to 9. True labels were averaged between self-assessment and two expert ratings. Spearman's cross-correlation coefficient, denoted as $\rho$, is used as the evaluation metric.

### 3.3.4 Systems

We compare our systems to the following baselines, from which we report the best-performing systems: (i) the ComParE 2013 feature set fed into an SVR, (ii) ComParE 2013 feature set quantized with BoAW with 500 centroids, (iii) S2SAE-embeddings with -70dB clipping fed into SVR, as well as a score-fusion of these three systems.

We used the raw waveform-based CNN framework as described in Section 3.1 with an input segment length $w_{seq}$ of 250ms. Even though sleepiness has 9 ordinal labels, we opted for a cross-entropy loss. Similar to the baseline system studies reported in (Schuller et al., 2019), we conducted studies with two experimental setups: (a) train the CNNs on the training data and test on development data and (b) train the CNNs on both training and development data and test on the test set. In each case, 5% of the data was used for cross-validation.

In addition to segmental and sub-segmental architecture, as introduced above, Table 3.5 shows the AF-CNN architecture, a sub-segmental model, that slightly differs in kernel widths and number of filters. The classification stage consists of one hidden layer with 100 units.

Table 3.5: CNN architectures for AF predictor. $N_f$: number of filters, $kW$: kernel width, $dW$: kernel shifts, $MP$: max-pooling.

| **Model** | Layer | Conv | | | MP |
|---|---|---|---|---|---|
| | | $N_f$ | $kW$ | $dW$ | |
| AF-CNN | 1 | 80 | 30 | 10 | 3 |
| | 2 | 60 | 7 | 1 | 3 |
| | 3 | 60 | 7 | 1 | 3 |

### 3.3.5 Integrating speech production knowledge

As discussed in Section 3.3.3, sleepiness can induce changes in the articulation process, i.e. in the speech production process resulting in slurred speech, and less crisp or incorrect pronunciation. In order to integrate articulatory information into our models, we investigated a transfer learning framework where the CNN is first trained to predict articulatory features (AFs) within four broad categories, namely, manner of articulation (e.g. degree of constriction), place (of constriction), height (of the tongue) and vowel. These AFs are inspired by recent work on articulatory feature-based speech recognition (Rasipuram and Magimai.-Doss, 2016). To predict the degree of sleepiness, we use the AF-initialized CNNs, replace the output layer with an output layer consisting of the nine sleepiness categories and train those models. Figure 3.2 summarizes this procedure. Knowledge from the 4 AF categories is utilized to initialize 4 separate CNNs, which are fine-tuned on the sleepiness data. We hypothesize that such an initialization helps to exploit articulatory differences due to sleepiness.

AF predictors are trained based on knowledge that maps phones to AFs. With such mapping, one can train acoustic-to-AF predictors by using an alignment of transcribed speech. The challenge data is not transcribed, so we used the AMI corpus (Carletta et al., 2005), which consists of 77 hours of speech. From this data, we used Kaldi to train HMMs for context-dependent phones, where the HMM states were jointly modeled by using subspace GMMs. The corresponding frame-to-phone alignments and the phone-to-AF mappings were then used to train the above-mentioned four AF-CNNs. The model architecture is similar to sub-segmental architecture and is described in Table 3.1 as AF-CNN, except in this case, the single hidden layer MLP contains 1024 hidden units.

We then adapted the resulting four AF-CNNs on the sleepiness data.

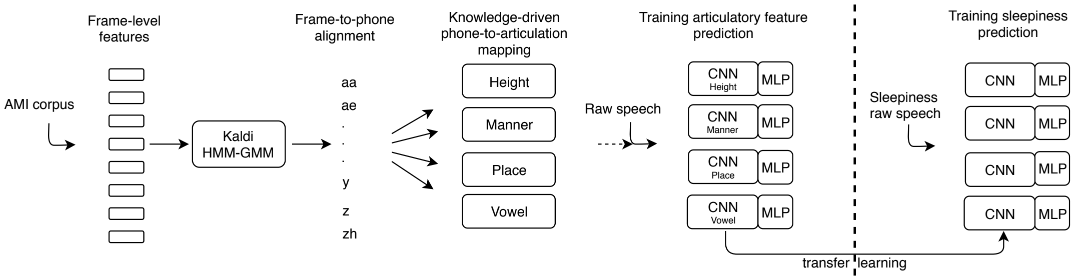


Figure 3.2: Overview of transfer learning for sleepiness prediction from CNNs that were initially trained to predict articulatory features.

**Posterior vector fusion with an MLP:** We also investigated combining different systems. For that, we used an MLP to fuse scores from different systems. The MLP had one hidden layer with 128 nodes with ReLU activation, a dropout layer with 10% and the output layer predicts the nine sleepiness categories.

### 3.3.6 Results

Table 3.6 compares the performance of the proposed systems with the baseline systems provided as part of the challenge and systems reported as part of the challenge. It is important to mention that the challenge allowed only five trials on the test set, hence only five test results for the proposed systems are reported.

On the first experimental setup i.e. training on the training set and evaluating on the development set, it can be observed that the proposed raw waveform modeling methods perform comparably to the best baseline systems and systems reported as part of the ComParE challenge.

Table 3.6: Results of all the presented CNNs on the ComParE 2019 sleepiness challenge data in Spearman's cross-correlation coefficient $\rho$. A + denotes a fusion using an MLP.

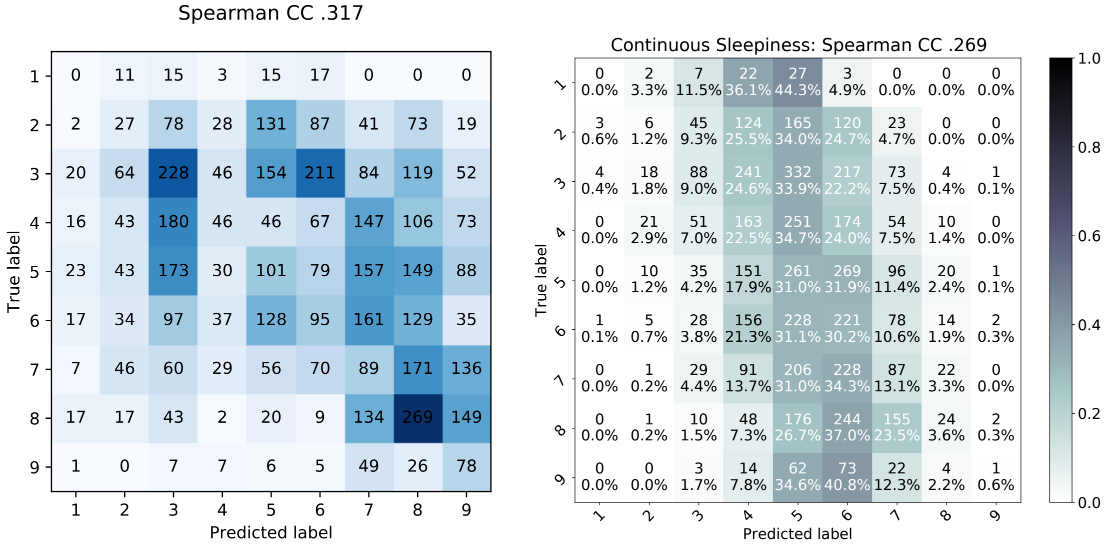
| **ComParE 2019 Baseline systems** | **Dev** | **Test** |
|---|---|---|
| $ComParE_{2013}$ Schuller et al. (2019) | .251 | .314 |
| $COMPARE_{2013}BoAW_{500}$ Schuller et al. (2019) | .250 | .304 |
| $S2SAE_{-70dB}$ Schuller et al. (2019) | .261 | .310 |
| 3-best Fusion Schuller et al. (2019) | - | .343 |
| **Competition systems** | | |
| Elsner et al. (2019) | .290 | .335 |
| Yeh et al. (2019) | .373 | .369 |
| Wu et al. (2019b) | .343 | - |
| Ravi et al. (2019) | .300 | .331 |
| Gosztolya (2019) | .367 | .383 |
| Wu et al. (2019a) | .326 | .365 |
| **Proposed raw waveform CNNs** | | |
| $subseg$ | .280 | .201 |
| $seg$ | .274 | .222 |
| **Proposed AF-CNNs** | | |
| $height$ | .267 | - |
| $manner$ | .292 | - |
| $place$ | .262 | - |
| $vowel$ | .295 | .312 |
| **Proposed fusion** | | |
| $manner + place + vowel$ | 304 | - |
| $manner + place$ | .311 | - |
| $manner + vowel$ | .317 | .325 |
| $manner + seg$ | .315 | - |
| $manner + vowel + seg$ | .319 | - |
| $manner + seg + ComParE$ | .329 | - |
| $manner + seg + BoAW_{500}$ | .344 | .321 |

We can observe that score fusion leads to improvement in performance. Thus, indicating that different CNNs are capturing complementary information. When comparing on the second experimental setup, i.e. training on train and development set and evaluating on the test set, we can see that the raw waveform CNNs not necessarily generalize well. However, the AF-CNN and fusion systems generalize well. This shows that integrating speech production knowledge is indeed aiding in predicting degree of sleepiness and yields comparable systems.

Besides the proposed systems, Elsner et al. (2019) and Wu et al. (2019b) investigated modeling raw waveform using CNNs for the sleepiness challenge. In (Wu et al., 2019b), a system based on CNN-BLSTM yielded significantly poor results. In (Elsner et al., 2019), it was found that a CNN-based system using a considerably longer window of speech input, more precisely 1.5 s speech, without data augmentation yielded a competitive system. In our case, the raw waveform-based CNNs without modeling speech production knowledge model 250 ms of speech at the input. This difference could possibly explain low performance on the test set. However, when integrating speech production knowledge, although the CNN hyperparameters were chosen from previous speech recognition studies, we can observe that with 250 ms speech input, we yield competitive systems. This suggests that raw waveform CNNs and AF-CNNs are modeling different information.

### 3.3.7 Analysis

We performed a confusion matrix analysis of the results obtained in the first experimental setup. Figure 3.3 in (a) shows the confusion matrix of our system *manner* + *vowel*, under (b) shows the system that performed best on the development set: BoAW$_2$000 (BoAW$_2$000 is not in Table 3.6 we picked systems based on best test performance). Unlike the baseline system (Schuller et al., 2019), it can be observed that classifications are spread over all degrees of sleepiness. We have the highest accuracy for KSS ratings of 3 and 8, meaning that our system is able to differentiate the extreme sleepiness categories well. In contrast, accuracy is lower for KSS ratings between 4 and 6, which are naturally difficult to distinguish. Moreover, the highest number of predictions are reasonably spread along the diagonal. KSS label 1 is not correctly classified, presumably because of a lack of samples – at least 5 times less in both training and development set than KSS labels 2 to 8. In general, we found similar trends in other systems that we investigated.



(a) *manner* + *vowel* (b) BoAW$_{2000}$

Figure 3.3: Confusion matrix of the score fusion from the CNNs *manner* and *vowel* as well as the baseline system BoAW$_{2000}$ (cf. Table 3.6).

To get an impression of what frequency regions the first convolutional layer is focusing on, we computed the cumulative frequency response (CFR) as follows (Palaz et al., 2019):

$$F_{cum} = \sum_{k=1}^{N_f} F_k / ||F_k||_2 \tag{3.1}$$

$N_f$ denotes the number of filters and $F_k$ is the frequency response of filter $f_k$. Figure 3.4 compares the CFR for raw waveform-based systems. In both *subseg* CNN and *seg* CNN frequency regions around 1000 Hz or below are given emphasis.

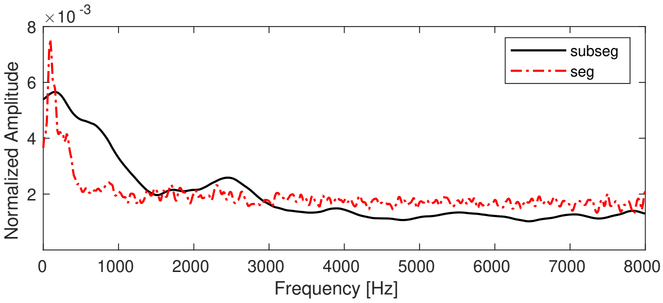


Figure 3.4: Cumulative frequency responses of first convolutional layer from raw waveform CNNs.

Figure 3.5 shows the CFR for AF-CNNs after adaptation/training on sleepiness challenge data to estimate the degree of sleepiness. It can be observed that there are differences in the information modeled by the CNNs for different AFs. However, in general, the emphasis on frequency regions is similar to CNNs trained for speech recognition task (Palaz et al., 2019). Furthermore, when compared to raw waveform CNNs (Figure 3.4), the CFRs are very different, i.e. emphasis is given to frequencies above 1000 Hz that are associated with the articulation aspect of speech. This indicates that indeed the raw waveform CNNs and AF-CNNs are focusing on different information. In addition, it also explains the performance gains obtained when fusing these systems.

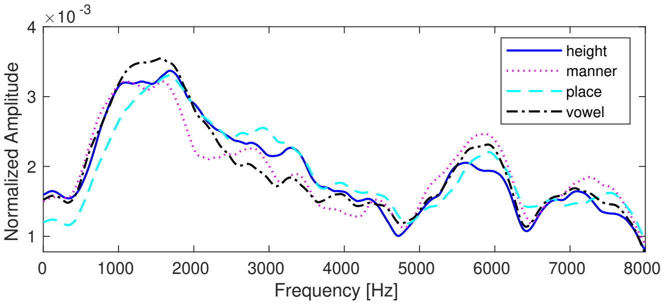


Figure 3.5: Cumulative frequency responses of first convolutional layer from AF-CNNs.

### 3.3.8 Conclusions

This section reviews different approaches for estimating the degree of sleepiness. Our investigations showed that integrating phonetic knowledge yields better systems when compared to simply modeling raw waveforms. Among the AF-CNNs, the manner CNN and vowel CNN yield the best systems. Especially a score fusion of the manner + vowel models, as well as manner+seg+BOAW$_{500}$, hence a fusion with a baseline system yield competitive performance. Our analysis of the first convolution layer shows that raw waveform CNNs and AF-CNNs focus on different frequency information, hence capture complementary information. This could be exploited through score fusion.

## 3.4 Summary

In this chapter, we explored how to induce knowledge into raw waveform convolutional neural networks. The following three main lines of results emerged: (i) Filtering signals was shown to let the CNN focus on specific aspects. Similarly, (ii) setting the kernel width of the CNN lets the model focus on either source or vocal tract information of speech. For Alzheimer's assessment, a combination of both neural and non-neural systems yielded particularly competitive results. Finally (iii) pre-training the CNN on prediction of articulatory features was successfully shown to be effective when fine-tuning on degree sleepiness estimation, which affects articulation. This unique pre-training method can be useful for tasks related to articulation in low-resource scenarios.

# 4 On modeling glottal source information for breathing impairment assessment in Parkinson’s disease

In this chapter, we propose a pilot study on breathing impairment assessment from the speech of Parkinson’s disease patients. Parkinson’s disease produces several motor symptoms, including different speech impairments that are known as hypokinetic dysarthria. Symptoms associated with dysarthria affect the main dimensions of speech such as articulation, prosody, intelligibility, and also phonation. Studies in the literature have mainly focused on the analysis of articulation and prosody because they seem to be the most prominent symptoms associated with dysarthria severity. However, phonation impairments also play a significant role in evaluating the global speech severity of Parkinson’s patients. Phonation is the process of passing air through the vocal folds to produce speech sounds through quasi-periodic vibration. In this chapter, the goal is to find methods that can model this signal and what features are suitable to represent it. Therefore, this chapter proposes an extensive comparison of different methods to automatically evaluate the severity of specific breathing impairments in Parkinson’s patients. The considered methods include: (i) the computation of hand-crafted features, (ii) CNN-based models trained in an end-to-end manner, (iii) neural embeddings from the CNN models that are in-domain, and (iv) neural embeddings from out-of-domain pre-trained models. Implicitly, we built on the previous Chapter 3, in which we applied CNN framework to model articulation and prosody tasks. In the following, we conduct experiments to automatically classify between speakers with low versus high breathing impairment severity due to the presence of dysarthria and also to evaluate the severity of the breathing impairments on a continuous scale, posed as a regression problem.

## 4.1 Problem statement & related works

Parkinson’s disease (PD) is a neurological disorder characterized by the progressive loss of dopaminergic neurons in the midbrain. It affects approximately 10 million people worldwide, with a doubling of the global burden over the past 25 years because of the increase in longevity of people thanks to modern medicine methods (Dorsey et al., 2018). PD produces different motor and non-motor symptoms in patients. Motor symptoms include tremor, slowed movement, rigidity, bradykinesia, and lack of coordination, among others. Approximately 70-90%

of PD patients develop a multidimensional speech impairment called hypokinetic dysarthria (Logemann et al., 1978; Ho et al., 1998), which manifests itself typically in the imprecise articulation of consonants and vowels, mono loudness, mono pitch, inappropriate silences and rushes of speech, dysrhythmia, reduced vocal loudness, and harsh or breathy vocal quality. All these symptoms affect the phonation, articulation, prosody, and intelligibility aspects of the speech of PD patients (Rusz et al., 2011; Knuijt et al., 2017; Vásquez-Correa et al., 2018).

Dysarthria severity is usually evaluated with perceptual scales such as the Frenchay dysarthria assessment (Enderby and Palmer, 2008), or the Radboud dysarthria assessment (Knuijt et al., 2017), which evaluate different speech dimensions such as phonation, articulation, prosody, resonance, among others. Different studies in the literature have focused on the automation of the evaluation process of these speech dimensions in order to assess the global dysarthria severity of patients. Most of those studies have mainly focused on the automatic analysis of articulation and prosody because they seem to be the most prominent symptoms associated with dysarthria severity. Articulation impairments have been modeled with speech features based on the vowel space area (Rusz et al., 2013), formant frequencies, voice onset time (Montaña et al., 2018), the energy content in onset transitions (Orozco-Arroyave, 2016), and recent models based on convolutional neural networks (CNNs) (Vásquez-Correa et al., 2017a) and posterior probabilities of certain phonemic classes (Cernak et al., 2017; Moro-Velazquez et al., 2019). Prosody deficits have been commonly evaluated with features related to pitch, intensity, and duration (Bocklet et al., 2013; Norel et al., 2020).

Despite the fact that articulation and prosody are the most studied speech dimensions in hypokinetic dysarthria, phonation impairments also play a significant role in evaluating the global speech severity of PD patients. Phonation symptoms are related to the stability and periodicity of the vocal fold vibration, and difficulties in the process of producing air in the lungs to make the vocal folds vibrate. Different phonation deficits appear in PD patients' speech, including differences in glottal noise compared to healthy speakers, incomplete vocal fold closure, and vocal folds bowing, which are typically characterized by measures such as noise-to-harmonics ratio, glottal-to-noise excitation ratio, and voice turbulent index, among others (Tanaka et al., 2011). Additional phonation features include perturbation measures such as jitter, shimmer, amplitude perturbation quotient (APQ), pitch perturbation quotient (PPQ), and nonlinear dynamics measures (Arias-Vergara et al., 2017; Travieso et al., 2017), as well as features extracted from the reconstruction of the glottal source signal such as the quasi-open quotient, the normalized amplitude quotient, and the harmonic richness factor (Kadiri and Alku, 2019; Novotnỳ et al., 2020; Narendra and Alku, 2020). However, it is not clear whether these traditional features are able to properly characterize specific phonatory impairments that appear in the speech of PD patients because they are usually only considered to classify PD vs. healthy control (HC) speakers.

## 4.2 Modeling glottal source information

The ground hypothesis for breathing parameter assessment through glottal source information is that the rate of vocal fold oscillations is largely determined by the subglottal pressure (Alku, 2011). The increased rigidity of vocal folds affects glottal signals; therefore, we propose to use filtering to obtain glottal signals and analyze them. Then, phonatory impairments in PD patients can be evaluated using different feature extraction strategies on raw waveforms as well as the following non-exhaustive selection of filtering methods.

### 4.2.1 Filtering methods

We considered the following methods to reconstruct the glottal source information:

1. Iterative and/or Adaptive Inverse Filtering (**IAIF**)(Alku, 1992) is based on linear prediction (LP) filters that are computed in a two-stage procedure. This method is based on iterative refinement of both the vocal tract and the glottal components. The glottal excitation is obtained by canceling the effects of the vocal tract and lip radiation by inverse filtering. A Matlab implementation of this method is available online [1].

2. Glottal closure/opening instant estimation forward-backward algorithm (**GEFBA**), which is based on detecting instants of significant excitation (epochs) for high-resolution glottal activity detection (Koutrouvelis et al., 2015). GEFBA estimates the instants of glottal closures for determining the boundaries of glottal activity by assuming that two consecutive voiced regions differ by a distance greater than twice the maximum pitch period. A Matlab implementation of this method is available online [2].

3. Zero frequency filtering (**ZFF**), as already described in Section 3.2.3, ZFF is designed for epoch extraction and aims to remove all the influence from the vocal tract system in the speech waveform.

Figure 4.1 shows the difference between the raw speech waveform, the IAIF and GEFBA methods used to reconstruct the glottal signal, and the ZFF signal. These four signals are used to evaluate the phonation impairments that appear in PD patients.

### 4.2.2 Perturbation features

Perturbation features are used to model abnormal patterns in the vocal fold vibrations. Perturbation features can be extracted from the raw speech waveforms and from the ZFF signals. The feature set includes seven descriptors: (1-2) *Jitter* and *shimmer* to describe temporal perturbations in the fundamental frequency and amplitude of the speech signal, respectively

[1] https://github.com/covarep/covarep/blob/master/glottalsource/iaif.m

[2] http://cas.tudelft.nl/data/richard/GEFBA.zip

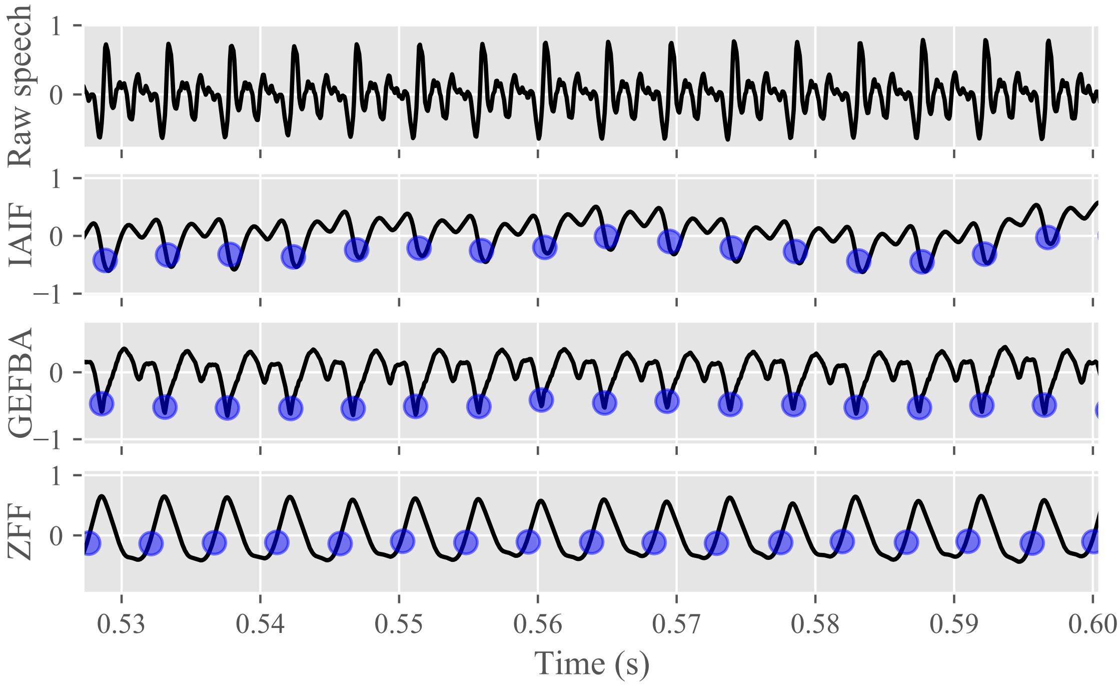


Figure 4.1: Different signals extracted from the phonation of a sustained vowel /ah/ are considered to evaluate the phonation impairments from PD patients. Blue dots indicate the detected GCIs.

Arias-Vergara et al. (2017). (3) APQ, which aims to measure the long-term variability of the peak-to-peak amplitude of the speech signal, by using a smoothing factor of 11 voiced periods. (4) PPQ to measure the long-term variability of the fundamental frequency, with a smoothing factor of five periods. (5-6) The first and second derivatives of the fundamental frequency contour, and (7) the log energy as a measure of loudness. Four statistical functionals are calculated per descriptor (mean, standard deviation, skewness, and kurtosis), forming a 28-dimensional feature vector per utterance.

### 4.2.3 Glottal features

Glottal features are computed over the reconstructed glottal signals using the IAIF and the GEFBA methods. Glottal features are focused on specific parts of the glottal cycle such as the opening and closing phases. The proposed feature vector comprises nine descriptors: (1) the temporal variability between consecutive GCIs, (2-3) the average and variance of the *Open Quotient (OQ)*, which is the ratio of the duration of the opening phase and the duration of the glottal cycle. (4-5) the average and variance of the *Normalized Amplitude Quotient (NAQ)*, which is defined as the ratio of the maximum of the glottal flow and the minimum of its derivative. (6-7) the average and variance of *H1H2*, which is the difference between the first two harmonics of the glottal flow signal. Finally (8-9) are the average and variance of the *Harmonic Richness Factor (HRF)*, which is the ratio of the sum of the harmonics amplitude and the amplitude of the fundamental frequency. These features are computed for every glottal cycle in segments with 200 ms length in order to measure short-term perturbations of the glottal flow. Finally, similar to the perturbation features, four statistical functionals are calculated per descriptor (mean, standard deviation, skewness, and kurtosis), forming a 36–dimensional feature vector per utterance.

The source code to extract the IAIF-based glottal signals and to compute the perturbation and glottal features is available online for the research community via the DisVoice toolkit[3].

## 4.3 Experimental study

We propose a pilot study to evaluate how to assess breathing impairment by modeling glottal source information. Towards that, we use propose different signals/filtering techniques, different feature extraction strategies, and classifiers/regressors for automatic evaluation. We perform our study on a subset of the PC-GITA corpus (cf. Section 4.3.2).

### 4.3.1 Proposed approach

Towards breathing impairment assessment in Parkinson's disease patients, we propose to compare hand-crafted features and raw-waveform CNN models. As hand-crafted features, we consider (i) perturbation features, that are extracted from either raw speech waveforms or the ZFF signals, and (ii) glottal features extracted from the IAIF and GEFBA reconstructed glottal signals. For end-to-end modeling, we propose raw waveform CNNs as introduced in Section 3.1, which can be trained on, the raw speech waveforms, IAIF signals, GEFBA signals, or ZFF signals.

We conduct our experiments in a 10-fold cross-validation strategy. All systems are trained to solve either the classification problem (low vs. high phonation impairments) or the regression problem (severity of the phonation impairment). All systems are applied to the three problems described in Section 4.3.2, namely breathing duration, breathing capacity, and global breathing impairment. The latter is the combination of the breathing duration and capacity scores.

### 4.3.2 PC-GITA corpus

The proposed systems are evaluated on the PC-GITA corpus (Orozco-Arroyave et al., 2014a). The data comprises utterances from 50 PD patients and 50 HC subjects, Colombian Spanish native speakers. The participants were asked to pronounce 10 sentences, six diadochokinetic (DDK) exercises, one text with 36 words, the sustained phonation of vowels, and a monologue. All patients were evaluated by a neurologist expert according to the MDS-UPDRS-III scale (Goetz et al., 2008), and they were recorded in ON state. The dysarthria severity of

[3] https://github.com/jcvasquezc/DisVoice

the participants was evaluated according to the m-FDA scale (Vásquez-Correa et al., 2018), which consists of 13 items and evaluates seven aspects of the speech including breathing, lips movement, palate/velum movement, laryngeal movement, intelligibility, and monotonicity. Each item ranges from 0 to 4 (integer values), thus the total score ranges from 0 (healthy speech) to 52 (completely dysarthric). Two items of the m-FDA scale are related to phonation impairments of the patients and include breathing duration (BD) and breathing capacity (BC) when participants pronounce sustained phonation of vowels and DDK tasks (cf. item 1 and 2 in Appendix B of (Vásquez-Correa et al., 2018)). The ratings of such items are used to evaluate the proposed models. We consider as well the global m-FDA breathing impairment score, which combines information about BD and BC (it ranges from 0 to 8). For this study, we only considered data from the phonations of sustained vowels and DDK tasks, which were the recordings used by the phoniatrician to label the phonation severity of the participants. Table 4.1 shows clinical and demographic information from the participants of this study.

Table 4.1: Demographic information from the participants in this study. **BD**: breathing duration, **BC**: breathing capacity.

| | **PD (n=50)** | **HC (n=50)** | **PD vs. HC** | **F vs. M** |
|---|---|---|---|---|
| **Sex (F/M)** | 25/25 | 25/25 | - | – |
| **Age** | 61.0 (9.3) | 61.0 (9.4) | 0.49$^a$ | 0.29 |
| **Years since diagnosis** | 10.6 (9.1) | - | - | – |
| **MDS-UPDRS-III** | 37.7 (18.1) | - | - | – |
| **MDS-UPDRS-speech** | 1.3 (0.8) | - | - | – |
| **Total m-FDA** | 28.8 (8.3) | 8.5 (7.4) | $\ll$0.005$^a$ | 0.28$^a$ |
| **m-FDA-BD** | 2.6 (1.0) | 1.0 (0.9) | $\ll$0.005$^a$ | 0.21$^a$ |
| **m-FDA-BD (high/low)** | 37/13 | 8/42 | $\ll$0.005$^b$ | 0.71$^b$ |
| **m-FDA-BC** | 2.5 (0.9) | 0.7 (0.7) | $\ll$0.005$^a$ | 0.25$^a$ |
| **m-FDA-BC (high/low)** | 37/13 | 2/48 | $\ll$0.005$^b$ | 0.12$^b$ |
| **m-FDA breath** | 5.1 (1.7) | 1.7 (1.4) | $\ll$0.005$^a$ | 0.18$^a$ |
| **m-FDA breath (high/low)** | 40/10 | 8/42 | $\ll$0.005$^b$ | 0.84$^b$ |

$^a$p-values calculated using Mann-Whitney U tests

$^b$p-values calculated using Chi-squared tests

The m-FDA labels for BD and BC are converted into high/low scores based on a threshold (median value of the scores assigned to the patients). Those subjects with scores lower than two are assigned low phonation severity. Conversely, subjects with an item higher or equal to two are labeled as patients with high phonation impairments. Hence, we decided to solve either a regression problem on the full range of the m-FDA sub-scores or a classification problem to evaluate low vs. high phonation impairment. The distribution between PD and HC subjects and the assigned m-FDA labels are gender-balanced (all p-values> 0.05) and age-balanced (Spearman's correlation between age and m-FDA scores are lower than 0.2 with all p-values> 0.05). Hence, the influence produced by demographic data on our problem can be discarded.

### 4.3.3 Systems

The extracted perturbation and glottal features were used to train an SVM classifier for the classification task and an SVR, both with a Gaussian kernel. The SVM/SVR hyperparameters $C$ and $\gamma$ were optimized in a randomized-search strategy (Bergstra and Bengio, 2012) based on the development set accuracy.

The 1-D CNN models are consistent with what is described in Section 3.1. We deploy an architecture consisting of four 1D convolutional layers, followed by a 10-dimensional hidden layer and an output layer. In order to guide the learning procedure, the first layer's filters' kernel length is relevant. As distinguished in Table 3.1, the sub-segmental (filter length < 1 pitch period) is applied in this work, since it tends to rather focus on source-related characteristics (Dubagunta et al., 2019). The classification task was trained with a binary cross-entropy loss function and a sigmoid function at the output; the regression task with mean-squared-error loss and a linear output function. In both cases, the starting learning rate is $1e-3$, which is halved after an epoch in which the validation loss did not reduce. Early stopping method was used to stop the training.

### 4.3.4 Results

All results are shown in Table 4.2. Note, that even though classification and regression results are presented in the same line, the systems are trained separately. The results obtained classifying high vs. low phonation impairments are shown in the left half of Table 4.2. In general, the best results are observed using perturbation features computed either from the raw speech waveform or from the ZFF signals. Regarding the two methods for glottal source estimation, higher accuracies are observed with the glottal signals computed using the GEFBA method. The accuracies obtained with the raw waveform CNNs are not as high as expected. However, note that moderate results are observed when the CNNs are trained with the ZFF signals.

The accuracy to assess breathing duration ranges from 56 to 79% depending on the considered method. The highest accuracy is obtained with the computation of perturbation features over the ZFF signals. Similar accuracies are observed with the raw speech waveform. The highest accuracy for the breathing capacity (84%) is obtained as well with the perturbation features, but in this case computed upon the raw speech waveform, followed by the perturbation features computed upon the ZFF signals. Finally, the accuracy for the assessment of global breathing impairments ranges from 54 to 76%. Similar accuracies are observed with the perturbation features computed upon the raw speech waveform and the ZFF signals.

The results about the continuous evaluation of the phonation impairments of the participants using a regression approach are presented in the right half of Table 4.2. The results are presented in terms of Pearson's correlation coefficient ($r$), Spearman's correlation coefficient ($\rho$), and mean absolute error (MAE). Strong correlations are obtained for the three addressed

Table 4.2: Results of classification and regression of different breathing impairments in PD patients.

| Signal | Features | ACC [%] | F-score | SENS [%] | SPEC [%] | r | $\rho$ | MAE |
|---|---|---|---|---|---|---|---|---|
| | | | | m-FDA Breathing duration | | | | |
| Raw | Perturbation | 78 | 77 | 80 | 76 | **.659** | **.662** | **0.86** |
| Raw | CNN | 65 | 52 | 47 | 80 | **.379** | **.427** | **1.12** |
| IAIF | Glottal | 71 | 70 | 62 | 78 | .436 | .444 | 1.00 |
| IAIF | CNN | 60 | 28 | 26 | 89 | .016 | -.034 | 1.46 |
| GEFBA | Glottal | 76 | 75 | 64 | 85 | .426 | .457 | 1.00 |
| GEFBA | CNN | 56 | 37 | 44 | 66 | .214 | .182 | 1.43 |
| ZFF | Perturbation | **79** | **78** | **73** | **84** | .591 | .603 | 1.00 |
| ZFF | CNN | **70** | **57** | **56** | **83** | .075 | .076 | 1.86 |
| | | | | m-FDA Breathing capacity | | | | |
| Raw | Perturbation | **84** | **83** | **77** | **89** | **.660** | **.659** | **0.86** |
| Raw | CNN | 65 | 43 | 35 | 83 | **.354** | **.315** | **1.31** |
| IAIF | Glottal | 65 | 62 | 51 | 74 | .308 | .429 | 1.00 |
| IAIF | CNN | 43 | 34 | 54 | 40 | .003 | -.039 | 1.44 |
| GEFBA | Glottal | 72 | 71 | 74 | 70 | .460 | .510 | 1.00 |
| GEFBA | CNN | 44 | 19 | 29 | 54 | -.121 | -.109 | 1.45 |
| ZFF | Perturbation | 80 | 79 | 79 | 80 | .659 | .683 | 0.89 |
| ZFF | CNN | **69** | **42** | **37** | **89** | .125 | .102 | 1.67 |
| | | | | m-FDA Global Breathing impairments | | | | |
| Raw | Perturbation | 76 | 75 | 69 | 83 | **.732** | **.741** | **1.40** |
| Raw | CNN | 56 | 53 | 56 | 55 | **.352** | **.341** | **1.27** |
| IAIF | Glottal | 60 | 59 | 48 | 71 | .129 | .474 | 2.00 |
| IAIF | CNN | 54 | 54 | 77 | 32 | .065 | .098 | 1.58 |
| GEFBA | Glottal | 71 | 70 | 65 | 77 | .528 | .620 | 1.91 |
| GEFBA | CNN | 49 | 47 | 57 | 43 | -.029 | .024 | 1.34 |
| ZFF | Perturbation | **76** | **76** | **77** | **75** | .673 | .714 | 1.54 |
| ZFF | CNN | **66** | **55** | **52** | **79** | .260 | .250 | 1.62 |

problems, especially using the perturbation features computed upon the raw speech waveforms and the ZFF signals. Similar to the classification results, the correlations observed with the raw waveform CNNs are not as high as expected; however, this can be explained by the little amount of data and the reduced variability of the labels to solve the regression problems. In addition, the results observed with the glottal signals estimated with the GEFBA method outperformed the ones obtained with the glottal signals estimated with the classic IAIF algorithm. Particularly, the best result is observed for the assessment of the global motor performance ($\rho$=0.741) probably because this is the scale with more variability in the labels (it ranges from 0 to 8), as compared to the breathing duration and breathing capacity, which only range from 0 to 4.

### 4.3.5 Analysis

In order to verify our initial hypothesis of estimating subglottal pressure through glottal signals, we are interested what the CNN's first convolutional layer filters are modeling. In Figure 4.2, we plotted the cumulative frequency response of the first layer filters of the models, as introduced in Equation 3.1 in Section 3.3.7. We plotted the CFR of the four models' global breathing impairment classification. We observe, that the raw and ZFF filters emphasize frequencies below 1kHz, which is desired. The frequency responses of IAIF and GEFBA on the hand have basically flat responses, so the models are not learning any specific structural information from the signals.

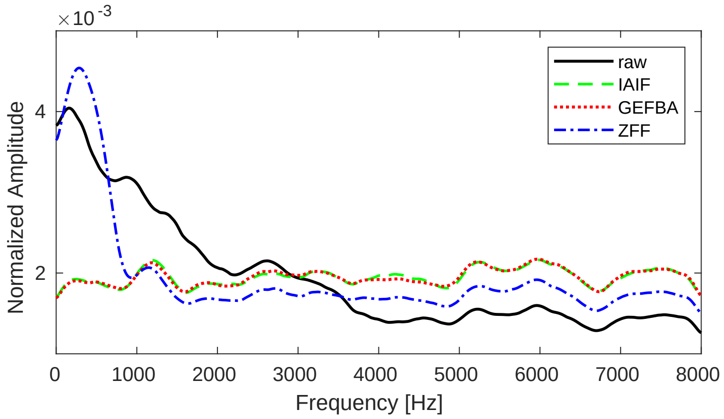


Figure 4.2: Cumulative frequency response of the first layer filters of the CNNs.

## 4.4 Neural embeddings for breathing parameter estimation

In the previous section, hand-crafted features outperform end-to-end CNN models. For a CNN-based assessment, we average the predictions of the CNNs into an utterance-level score. However, these systems could lack statistical information. Therefore, we investigated how statistical functionals of neural embeddings perform on breathing impairment assessment.

### 4.4.1 Proposed approach

We propose to use statistical functionals of neural embeddings as features for breathing parameter estimation. End-to-end modeling might lack statistical information, since output scores are averaged, meaning the decisions are averaged. However, embeddings, that are extracted before the decision-making may contain more relevant information. Therefore, we propose two kinds of neural embeddings: (i) task-dependent neural embeddings from raw-waveform CNN that are trained on the respective task from raw and ZFF signals, selected based on their good performance. In addition, we deploy embeddings from a CNN-based model that is trained to predict the actual breathing signal, which is referred to as UCL-SBM, and (ii) task-independent neural embeddings from general-purpose pre-trained models. We propose to use

VGGish embeddings and Wav2vec2.0 embeddings. From every embedding representation, the same four statistical functionals as for the hand-crafted features are calculated per descriptor: mean, standard deviation, skewness, and kurtosis. The resulting feature vector is then fed into an SVM/SVR for classification/regression. The experimental setup is unchanged: We perform 10-fold cross-validation.

### 4.4.2 Systems

The following embeddings are input to an SVM for classification and to an SVR for regression. In both cases, we use a radial basis function kernel and choose hyperparameter C based on the best validation performance.

**Raw** Embeddings from the 1-D CNN model trained on raw waveforms are extracted from the 10-dimensional last hidden layer before the activation. When calculating 4 functionals per dimension, a 40-dimensional feature vector per utterance is formed. Since convergence in the training of the CNN models can vary throughout the folds, we opted to pick embeddings from the fold with the best metrics on the validation set.

**ZFF**: Likewise, from the 1-D CNN model trained on ZFF signals we extracted 10- are extracted from the 10-dimensional last hidden layer before the activation to form a 40-dimensional feature vector per utterance. Since convergence in the training of the CNN models can vary throughout the folds, we opted to pick embeddings from the fold with the best metrics on the validation set.

**UCL-SBM:** In Nallanthighal et al. (2021), the authors train a CNN-based model to predict the actual breathing signal based on the UCL-SBM dataset (Schuller et al., 2020), which consists of free speech data and the corresponding breathing signals measured with a chest-sensor. The dataset contains about 3 hours of recordings. Notably, the architecture of this model is the exact same is for our own CNN-models (cf. Table 3.1). Input to the model are overlapping 3-second chunks of raw speech waveforms as input; a mean squared error loss is applied. From this model, we extract the 10-dimensional last hidden layer embeddings on a per-frame basis to form a 40-dimensional feature vector per utterance.

**VGGish:** is an embedding representation, trained on the popular VGG-16 architecture, that consists of 13 CNN and 3 dense layers. It is trained on the AudioSet dataset for acoustic scene classification. This dataset has 5.24 million hours of data, and a cross-entropy loss function is used. We extract 128-dimensional vectors per 1-second inputs of log mel spectrograms from a bottleneck layer right before the output (Hershey et al., 2017). In a post-processing step, a PCA transformation is applied, and the embeddings are quantized to 8-bit [4].

**Wav2vec2.0:** is a self-supervised pre-trained model for speech recognition (Baevski et al., 2020). It consists of filter stage of 5 CNN layers followed by a transformer architecture. We extract

[4] https://github.com/tensorflow/models/tree/master/research/audioset/vggish

embeddings from the wav2vec2.0 base model that is trained 53k hours of unlabeled data and fine-tuned on 960 hours of Librispeech. We extract 768-dimensional frame-level embeddings of the 6th transformer layer (out of 13), thus from the middle. Wav2vec2.0 embeddings have become a popular choice for paralinguistic tasks, as for example assessed in the SUPERB benchmark, where embeddings from self-supervised models dominate the leaderboard in 2022 (Yang et al., 2021).

### 4.4.3 Results

Table 4.3 shows our results: On the classification tasks, we observe that the task-dependent embeddings (raw, ZFF, and UCL-SBM) outperform the task-independent embeddings. The best results are obtained with the raw embedding, where we achieve an improvement of 14% absolute when classifying breathing duration, 13% when classifying breathing capacity, and 17% on the global breathing impairment compared to the end-to-end modeling. However, these improvements could not be translated into improvements in the regression task. On the regression task, the best embedding is the UCL-SBM embedding, which however still falls short of the hand-crafted features. Notably, wav2vec2.0 embeddings fail completely even though successful on other tasks.

Table 4.3: Results on classification and regression of breathing impairments in PD patients with statistical functionals of neural embeddings of dimension $Dim$.

| Embedding | $Dim$ | ACC [%] | F-score | SENS [%] | SPEC [%] | r | $\rho$ | MAE |
|---|---|---|---|---|---|---|---|---|
| m-FDA Breathing duration | | | | | | | | |
| **Raw** | 10 | 84 | 82 | 84 | 84 | .177 | .171 | 1.36 |
| ZFF | 10 | 69 | 64 | 61 | 76 | .111 | .113 | 1.39 |
| UCL-SBM | 10 | 64 | 61 | 67 | 61 | .368 | .361 | 1.22 |
| VGGish | 128 | 63 | 54 | 51 | 73 | .277 | .283 | 1.34 |
| Wav2vec2.0$_{mid}$ | 768 | 55 | 0 | 0 | 100 | .003 | .002 | 1.52 |
| m-FDA Breathing capacity | | | | | | | | |
| **Raw** | 10 | 82 | 74 | 71 | 89 | .153 | .148 | 1.40 |
| ZFF | 10 | 70 | 54 | 49 | 83 | .062 | .067 | 1.43 |
| UCL-SBM | 10 | 65 | 46 | 40 | 83 | .400 | .393 | 1.32 |
| VGGish | 128 | 65 | 46 | 41 | 82 | .356 | 0.37 | 1.29 |
| Wav2vec2.0$_{mid}$ | 768 | 61 | 0 | 0 | 100 | .001 | .002 | 1.59 |
| m-FDA Global Breathing impairments | | | | | | | | |
| **Raw** | 10 | 83 | 82 | .81 | .84 | .211 | .193 | 1.31 |
| ZFF | 10 | 68 | 66 | 64 | 73 | .111 | .102 | 1.35 |
| UCL-SBM | 10 | 64 | 65 | 71 | 58 | .373 | .369 | 1.23 |
| VGGish | 128 | 64 | 60 | 56 | 72 | .370 | .376 | 1.20 |
| Wav2vec2.0$_{mid}$ | 768 | 52 | 0 | 0 | 100 | .003 | .005 | 1.57 |

## 4.5 Summary

We performed a pilot study to evaluate the severity of different breathing impairments that appear in PD patients due to the presence of hypokinetic dysarthria. Our initial hypothesis was to filter signals to obtain glottal signals and analyze them. In our results, we show that this is not necessary, since both hand-crafted feature-based systems and CNN-based systems performed mostly best on raw waveforms, closely followed by ZFF-signal-based systems.

Additionally, we studied different embedding representations and showed that task-dependent embeddings outperform task-independent embeddings. We also showed that the advantage of feature-based systems over CNN-based systems could be compensated by computing more statistics from embeddings. CNN-based systems might suffer from the lack of data, but might be more robust to noise. When investigating the performance of pre-trained embeddings, we found that embeddings from a system trained to predict breathing signals worked well, since the task is very similar, but would probably benefit from fine-tuning. Task-independent off-the-shelf embeddings fall behind, notably trained on orders of magnitude more data, but may require fine-tuning on in-domain data.

Overall, we were able to discriminate between low vs. high breathing impairments with up to 83% accuracy and global breathing impairment regression with a Pearson's correlation of up to .732. The global breathing impairment rating is a combination of two ratings: Breathing duration impairment, which we demonstrated to solve in a classification with up to 82% accuracy, and regression with a Pearson's correlation of up to .659. Breathing capacity impairment classification was solved with up to 82% accuracy, regression with a Pearson's correlation of up to .660.

# 5 Intelligibility estimation: a divergence from healthy speech

Intelligibility is the ability of a listener to understand the message conveyed by a speaker. Intelligibility is a subjective measure, which is usually assessed by listening tests. In this chapter, we propose a novel approach to estimating the intelligibility of atypical speakers. We propose to emulate subjective listening tests: Typically, listeners transcribe a recording or compare it to a prompt. Based on that, the uttered speech recording can be classified as correct or incorrect. Clinicians aggregate multiple such decisions, e.g. for severity categorization. A speaker's intelligibility is estimated as the percentage of correct words uttered, averaged over the decisions of one or multiple listeners. In this chapter, for atypical intelligibility assessment, we propose to compare words to the same word uttered by control speakers. We replace a listener with a speaker and thereby take advantage of the production-perception feedback loop. Implicitly, we measure if or by how much a recorded word deviates from what is considered normal or healthy speech, i.e. a divergence. Along this paradigm, we investigated several variations in terms of the speech signal representation and the amount of healthy speech used. In this chapter, we demonstrate this approach on a corpus of dysarthric speech in Section 5.3 and on a small sample of lip filler surgery recordings in Section 5.4 to confirm the results of a listening test.

## 5.1 Related work

Previous work on objective dysarthric speech intelligibility assessment can be broadly grouped into two categories:

i) assessment without explicit use of linguistic information: Legendre et al. proposed prediction of intelligibility using amplitude modulation spectra (Legendre et al., 2009). Falk et al. (2011) investigated modeling of short- and long-term temporal dynamics information. In (Falk et al., 2012), a signal processing-based composite measure was proposed, inspired by the notion that intelligibility can be expressed as a linear combination of perceptual dimensions phonation, nasality, articulation and prosody (De Bodt et al., 2002). Janbakshi et al. proposed the P-ESTOI measure (Janbakhshi et al., 2019a), which builds upon the speech intelligibility measures short-time objective intelligibility (STOI) (Taal et al., 2011) and extended-STOI (Jensen and Taal, 2016). Different subspace-based methods such as iVector-based (Martínez et al., 2015), use of spectral subspaces extracted through principal component analysis or approximate joint diagonalization (Janbakhshi et al., 2019b) have also been proposed. The subspace methods assess intelligibility by measuring the deviation or distance between the control speech and dysarthric speech in the trained subspace.

ii) assessment based on explicit use of linguistic information: Kim et al. (2015) proposed an approach where automatic speech recognition (ASR) with a confusion network is used to obtain "phone-to canonical-phone" mappings. These mappings are summarized in per-speaker histograms for a defined set of words and are then used to estimate an intelligibility score for each speaker. Middag et al. (2009) proposed an approach where the dysarthric speech is aligned using an ASR system to obtain phone probabilities or phonological feature probabilities based confidences. These confidences are then accumulated over a specified group of phones for each speaker to estimate an intelligibility score. Finally, ASR system accuracy-based intelligibility assessment has also been investigated (Ferrier et al., 1995; Martínez et al., 2015).

In recent years, phone posterior feature-based speech assessment approaches have emerged, where sequences of phone posterior probabilities obtained from reference speech and test speech are matched for (a) speech codec and transmitted speech intelligibility assessment (Ullmann et al., 2015a), (b) synthesized speech intelligibility assessment (Ullmann et al., 2015a), and (c) degree of nativeness assessment (Rasipuram et al., 2015). We took inspiration from these works.

## 5.2 Proposed utterance verification-based speech intelligibility assessment

In a clinical setting, speech intelligibility can be assessed through an isolated word pronunciation test, where a speaker pronounces a set of isolated words, and the speech intelligibility is measured as a percentage of correctly identified words by human listeners (Duffy, 2012; Kent et al., 1989). We propose to emulate this listening test by comparing a speaker's test utterance to control utterances and then perform utterance verification, i.e. deciding if it was pronounced correctly. Towards that, we need to measure the similarity (Section 5.2.1) between two utterances and then make a decision as described in Section 5.2.2. As a representation, we opt for phonetic posterior features, which are known for being speaker- and noise-invariant.

### 5.2.1 Sequence matching with dynamic time warping

The similarity of two utterances can be obtained by matching sequences of feature representations with dynamic time warping (DTW) (Sakoe and Chiba, 1978). The match between two sequences of features $\mathbf{Z} = (\mathbf{z}_1, \cdots \mathbf{z}_n, \cdots \mathbf{z}_N)$ and $\mathbf{Y} = (\mathbf{y}_1, \cdots \mathbf{y}_m, \cdots \mathbf{y}_M)$, where $N$ denotes the number of frames in sequence $\mathbf{Z}$ and $M$ denotes the number of frames in $\mathbf{Y}$. The dynamic programming recursion is as follows:

$$L(m,n) = l(\mathbf{y}_m,\mathbf{z}_n) + \min[L(m-1,n),\\ L(m,n-1), L(m-1,n-1)]\,, \tag{5.1}$$

where, $l(\mathbf{y}_m,\mathbf{z}_n)$ is the local match score. We conduct the majority of our experiments using the symmetric Kullback-Leibler divergence between $\mathbf{y}_m$ and $\mathbf{z}_n$,

$$l(\mathbf{y}_m,\mathbf{z}_n) = \frac{1}{2}\cdot[\sum_{d=1}^{D} y_{m,d}\log\frac{y_{m,d}}{z_{n,d}} + \sum_{d=1}^{D} z_{n,d}\log\frac{z_{n,d}}{y_{m,d}}], \tag{5.2}$$

where $D$ is the dimension of the feature vector $\mathbf{Y}$ and $\mathbf{Z}$. $L(m,n)$ is the accumulated match score at $(m,n)$. The dynamic programming results in a global match score $L(M_k,N)$, which is then *normalized by the path length.* Figure 5.1 illustrates DTW for ease of imagination.

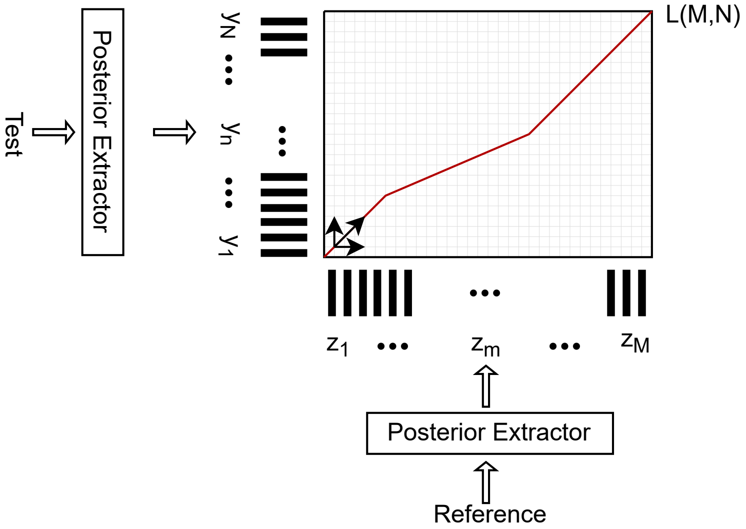


Figure 5.1: Illustration of dynamic time warping (DTW) between two sequences $\mathbf{Z}$ and $\mathbf{Y}$ to the global match score $L(M,N)$.

In the literature, it is well known that comparison of probability distributions using KL-divergence and other measures such as Bhattacharya distance is equivalent to hypothesis testing and yields an estimate of log-likelihood ratio (Kailath, 1967; Blahut, 1974). The global match score $L(M,N)$ is a sum of KL divergence valuesbetween posterior probability distributions on the best matching path normalized by the path length. So, $L(M,N)$ can be interpreted as an estimate of the log-likelihood ratio of the test utterance being the same as the reference utterance, through which utterance verification can be carried out. This notion is further exploited in Section 5.3.4.

### 5.2.2 Utterance verification based on DTW match score

It can be argued that when speech is unintelligible, the uttered word tends to map to a word other than the target word. As a result, the listeners are not able to identify the target word. This could be formulated as an utterance verification problem, i.e. testing the hypothesis whether the speech utterances **Y** and **Z** correspond to the same word or not. A similar understanding has been recently applied to assess the intelligibility of text-to-speech synthesis systems (Ullmann et al., 2015b).

In order to decide if an utterance is pronounced correctly, we need to apply a threshold on $L(M,N)$. As illustrated in Figure 5.2, the threshold is determined in the following manner:

1. Creating same word utterance pairs from the control speakers' data, matching them, and obtaining a distribution of global match score for the same word hypothesis;
2. Creating different word utterance pairs from the control speakers' data, matching them, and obtaining a distribution of global match score for NOT the same word hypothesis; and
3. Determining the threshold at the intersection of the two distributions, referred to as $Thr_{inter}$ or at the center of the two means of the histogram, referred to as $Thr_{cen}$.

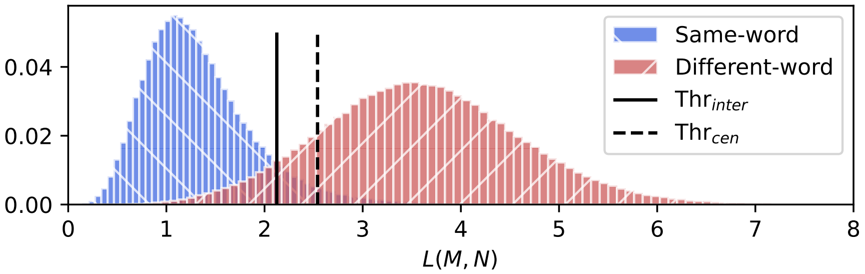


Figure 5.2: Distribution of same and different-word pair scores $L(M,N)$

**Intelligibility estimation:** Every speaker's intelligibility is estimated as the percentage of words that are identified as correct through utterance verification.

## 5.3 Dysarthric speech intelligibility assessment

To validate the proposed utterance verification-based speech intelligibility assessment method on a dysarthric speech database: the UA-Speech corpus. We propose to match utterances in posterior feature space. A set of experiments will compare different representations as well as the two introduced thresholds $Thr_{inter}$ and $Thr_{cen}$. To comply with the experimental setup in the literature, the utterance verification thresholds $Thr_{inter}$ and $Thr_{cen}$ are obtained using only recordings from the control speakers.

### 5.3.1 Posterior feature estimators

Speech representations vary greatly depending on the application. For dysarthric intelligibility assessment, we require a representation that describes the acoustic-phonetic production of a speech sound. In this work, we use posterior features, which are probability distributions over the phonetic space. In order to investigate the suitability of different representations, we investigated different posterior feature estimators. Broadly, we differentiate between **phone posterior space** and **broad phonetic or articulatory feature (AF) space**. Common to these representation spaces are i) They use probability distributions, which means divergence-based distance metrics are most suitable. ii) They are predicting a realization of speech sounds and therefore ideally should be noise-invariant as well as speaker-invariant. Typically, these estimators are trained on amounts of data larger than what is available in a clinical setting, we therefore use estimators trained on auxiliary data, which may be considered a convenience.

**Phone posterior space**: This representation is commonly used as features in speech recognition, and is trained from a frame-to-phoneme alignment. Our representation consists of 45-dimensional context-independent phoneme posterior probabilities estimated by an off-the-shelf multilayer perceptron (MLP). The MLP takes as input 39-dimensional perceptual linear predictive cepstral features with a frame size of 25 ms, a frameshift 10 ms, and a nine-frame temporal context (i.e. four frames preceding and four frames following). The MLP had a single hidden layer with 5000 units. The output layer consisted of 44 English phonemes (based on the UniSyn dictionary) and silence, i.e., $D = 45$. The MLP has been trained on 232 hours of conversational telephone speech (8kHz) with the QuickNet tool (Johnson et al., 2004) by minimizing the frame-level cross-entropy.

**AF space**: There are different ways to represent phonemes as articulatory features, e.g. as binary features (Chomsky and Halle, 1968) or multi-valued features (Ladefoged, 1993). Similar to phone posterior features, they are trained from a frame-to-phoneme alignment. However, instead of predicting phonemes, a mapping from phones to AF is used as targets of the predictor. We conducted studies with both binary features and multi-valued AF representations:

**AF**$_{binary}$ space consists of 18 binary valued AF predictors, namely for, {*pause, consonantal back, anterior, open, close, nasal, stop, continuant, lateral, flap, trill, voice, strident, labial, dental, velar, vocalic* }. In the Phonet toolkit[1] (Vásquez-Correa et al., 2019), these AFs are modeled by 18 off-the-shelf recurrent neural networks (RNN) based binary classifiers, i.e. $D = 18 \times 2$. The RNNs take as input log energies of 33-dimensional Mel filterbank energies. The RNN classifiers have been trained on 17 hours of clean FM podcasts in Mexican Spanish with a cost function based on cross-entropy. For more details, related to the mapping between Spanish phones and the AFs and training of RNNs can be found in (Vásquez-Correa et al., 2019).

**AF**$_{multi-manner}$ space consists of nine "manner of articulation" category AFs, namely, {*silence,*

[1] https://github.com/jcvasquezc/phonet

*aspirated, approximant, nasal, voiced-fricative, voiced-stop, stop, vowel*}. These AFs were modeled by an off-the-shelf convolutional neural network (CNN) that takes raw waveform as input and predicts the posterior probabilities of the nine manner of articulation category AFs, i.e. $D = 9$. The CNN has been trained on the 77 hour AMI corpus (Carletta et al., 2005) with a cost function based on cross-entropy. The mapping between the English phones and the AFs was based on previous work on automatic speech recognition (Rasipuram and Magimai.-Doss, 2016). For further details about the architecture and training of the CNN, the reader is referred to Fritsch et al. (2020).

### 5.3.2 UA-Speech corpus

The Universal Access (UA)-Speech database (Kim et al., 2008) consists of 15 English speakers with cerebral palsy (11 males, 4 females) and 13 healthy speakers (9 males, 4 females). Each impaired and control speaker has uttered 765 isolated words in total: 155 isolated words repeated 3 times and 300 isolated words spoken only once. In total, all cerebral palsy patients provided about 3 hours of speech; the control speakers were 1 hour and 45 minutes. In the database, each subject's intelligibility score has been obtained by having five naive listeners (native speakers of American English) transcribe the isolated words and then calculate the average number of correct transcriptions. The subjective intelligibility scores of the patients range from 2% to 95%. Similar to the previous works (Janbakhshi et al., 2019a,b), we use the 5th channel recordings for our experiments. An energy-based voice activity detection using Praat (Boersma, 2001) was used to extract the speech segments.

### 5.3.3 Results and analysis

Among the related works mentioned in Section 5.1, some were carried out on the UA-Speech database, and serve as baselines for our work. Table 5.1 shows Pearson's correlation ($r$) and Spearman's correlation ($\rho$) between subjective and objective intelligibility on the UA-Speech database.

Table 5.1: Related works on the UA-Speech database. Pearson's correlation ($r$) and Spearman's correlation ($\rho$) between subjective and objective intelligibility of the speakers with dysarthria only.

| | $r$ | $\rho$ |
|---|---|---|
| P – ESTOI (Janbakhshi et al. (2019a)) | .94 | .94 |
| Composite measure (Falk et al. (2012)) | .94 | .89 |
| iVectors (Martínez et al. (2015)) | .91 | - |
| Discriminant analysis (Paja and Falk (2012)) | .92 | |
| Spectral subspace (Janbakhshi et al. (2019b)) | -.83 | -.88 |
| Temporal dynamics (Falk et al. (2011)) | .87 | .85 |
| Word accuracy – based (Martínez et al. (2015)) | .89 | - |

Table 5.2 shows the results where an $IntScore$ was estimated for all 15 dysarthric speakers in the database. Under each of the correlation values, a $p$-value testing the hypothesis that the two sets of data are uncorrelated is also provided ($p$-values testing for a hypothesis test whose null hypothesis is that two sets of data are uncorrelated[1]).

It can be observed that the proposed approach consistently yields high Pearson's and Spearman's correlation coefficients for all the posterior feature spaces. Also, all the results are statistically significant. It is interesting to note that the choice of threshold does not influence the proposed approach's performance. Furthermore, the proposed approach consistently performs comparably to or better than the baseline approaches. As a point of reference regarding performance, we refer to Table 5.1.

Table 5.2: Pearson's correlation ($r$) and Spearman's correlation ($\rho$) between subjective and objective intelligibility at thresholds $Thr_{cen}$ and $Thr_{inter}$. $p$-values are presented in Italics font.

| Posterior feature space | $Thr_{cen}$ | | $Thr_{inter}$ | |
|---|---|---|---|---|
| | $r$ | $\rho$ | $r$ | $\rho$ |
| Phone | .939 | .939 | **.950** | **.957** |
| | *3.94e-7* | *2.31e-7* | *5.52e-8* | *2.29e-8* |
| $\text{AF}_{\text{binary}}$ | .918 | .885 | .922 | .885 |
| | *1.88e-6* | *1.13e-5* | *1.27e-6* | *1.32e-5* |
| $\text{AF}_{\text{multi-manner}}$ | .922 | .910 | .917 | .894 |
| | *1.01e-6* | *2.42e-6* | *1.43e-6* | *6.82e-6* |

In the above results, all 13 control speakers are used as references for a test utterance. However, less might be necessary, a lower number of control recordings would even be a desirable feature. Hence, we test the influence of **reducing the number of references** by randomly selecting $K$ control speakers per each test utterance. In the same vein, we propose to replace healthy recordings with **synthetic speech**. Given today's advances, it is possible to generate synthetic speech at a high degree of naturalness. We used an off-the-shelf neural TTS system Tacotron2 (Shen et al., 2018). The synthesizer has been originally trained on the LJSpeech corpus[2], which is an annotated English corpus including 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books. The system has been rated with a mean opinion score of $4.526 \pm 0.066$ on a scale of 1 to 5. During synthesis, each word from the UA-Speech word list was converted into a phoneme sequence based on CMUDict[3]. For more information about the TTS system, the reader is referred to (Shen et al., 2018). In Table 5.3, we present the results for the proposed approach using a single synthetic control speaker. It can be observed that the performance is comparable to the results obtained using the 13 control speakers.

Figure 5.3 summarizes the ideas presented above, showing the Pearson's and Spearman's corre-

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

[2] https://keithito.com/LJ-Speech-Dataset/

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Table 5.3: Pearson's correlation ($r$) and Spearman's correlation ($\rho$) between subjective and objective intelligibility from a single-synthetic-control. $p$-values are presented in Italics font.

| Posterior feature space | $Thr_{cen}$ | | $Thr_{inter}$ | |
|---|---|---|---|---|
| | $r$ | $\rho$ | $r$ | $\rho$ |
| Phone | .924 | .942 | .931 | **.961** |
| | *1.44e-7* | *8.08e-7* | *1.14e-8* | *4.46e-7* |
| $\text{AF}_{\text{binary}}$ | 0.827 | .893 | .822 | .885 |
| | *1.40e-4* | *7.23e-6* | *1.68e-4* | *1.13e-5* |
| $\text{AF}_{\text{multi-manner}}$ | **.937** | .906 | .930 | .912 |
| | *2.40e-7* | *3.09e-6* | *4.78e-7* | *2.13e-6* |

lation when the number of control speakers $K$ is reduced from 13 to 1 (selected randomly), as well as the results with a single synthetic reference. It can be observed that the performance is pretty stable when $K$ is reduced, even when selecting one single control speaker for intelligibility assessment, except for $\text{AF}_{multi-manner}$. This indicates that, in the proposed approach, the number of control speakers can be reduced considerably, without affecting the performance.

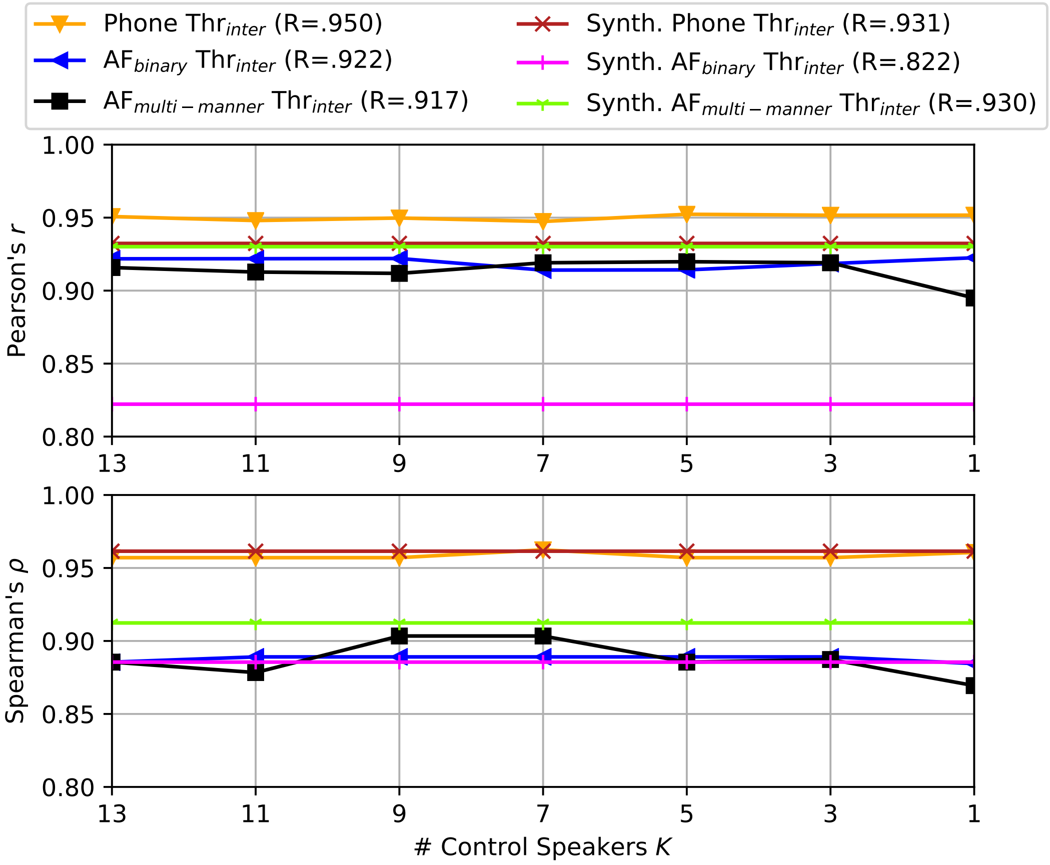


Figure 5.3: Pearson's correlation and Spearman's correlation when the number of control speakers $K$ is varied from 13 to 1. Synth refers to the case of single synthetic reference.

The proposed approach estimates an intelligibility score $IntScore$, i.e. percentage of words correct for each speaker with dysarthria, which can be directly related to the subjective listening score, without any intermediary mapping or regression. Fig. 5.4 shows the Pearson's correlation plot overlaid for the different systems, along with root-mean-square error (RMSE) between listener percentage word accuracy and the $IntScore$ (presented in the legends); each marker represents one speaker. It can be observed that phone space and AF$_{multi-manner}$ space are predicting well high intelligibility regions, while AF$_{binary}$ is predicting comparatively well the low intelligibility regions. As a consequence, although AF$_{binary}$ is not the best in terms of $r$ and $\rho$, it yields the best RMSE of 16.9%. We observe this trend even in the case of synthetic control speech, denoted as Synth AF$_{binary}$. This is promising as we have not used any dysarthric speech data to build any part of the assessment system. In the previous studies, on the same data set, RMSE ranging from 12% to 18.6% have been reported with the use of dysarthric speech data to build the intelligibility prediction models (Falk et al., 2012; Martínez et al., 2015). Overall, the analysis indicates that $IntScore$ estimation needs to be further improved for low intelligibility regions to take advantage of its interpretability. This is a part of our ongoing work.

### 5.3.4 Probabilistic utterance verification based on DTW match score

In Section 5.2.2, we propose atypical utterance verification by determining a threshold from in-domain data (or synthetic speech). In order to further reduce the necessary amount of healthy/in-domain speech, we propose a probabilistic approach: As mentioned in Section 5.2.1, the global match score $L(M,N)$ is a sum of KL-divergence-values along the best path and can be interpreted as an estimate of the log-likelihood ratio of the test utterance being the same as the reference utterance. Thus, the log-likelihood ratio estimate can be converted into a posterior probability of the test utterance being the same as the reference utterance by applying a logistic function:

$$P_c = \frac{L}{1+e^{-k(x-x_0)}} \tag{5.3}$$

where $L$ is the supremum, $x$ is the log-likelihood ratio $L(M,N)$, $x_0$ is the function's midpoint and $k$ the growth rate. In practice, we set $L = 1$, $x_0 \approx 1.2$ is approximately the middle of the same-word distribution in Figure 5.2 and serves to move the y-interception point to the positive side, and $k = 1$. Figure 5.5 shows the logistic function used to convert $L(M,N)$ into a posterior probability.

Finally, for atypical utterance verification, a posterior probability of a test utterance is considered correctly pronounced when $P_c \leq .5$ and rejected otherwise. Consistent with the previous experiments on the UA-Speech dataset, every utterance is matched against the same word uttered by the 13 control speakers, the resulting posterior probabilities are averaged, and the average posterior probability is used as the final verification score. Table 5.4 shows the

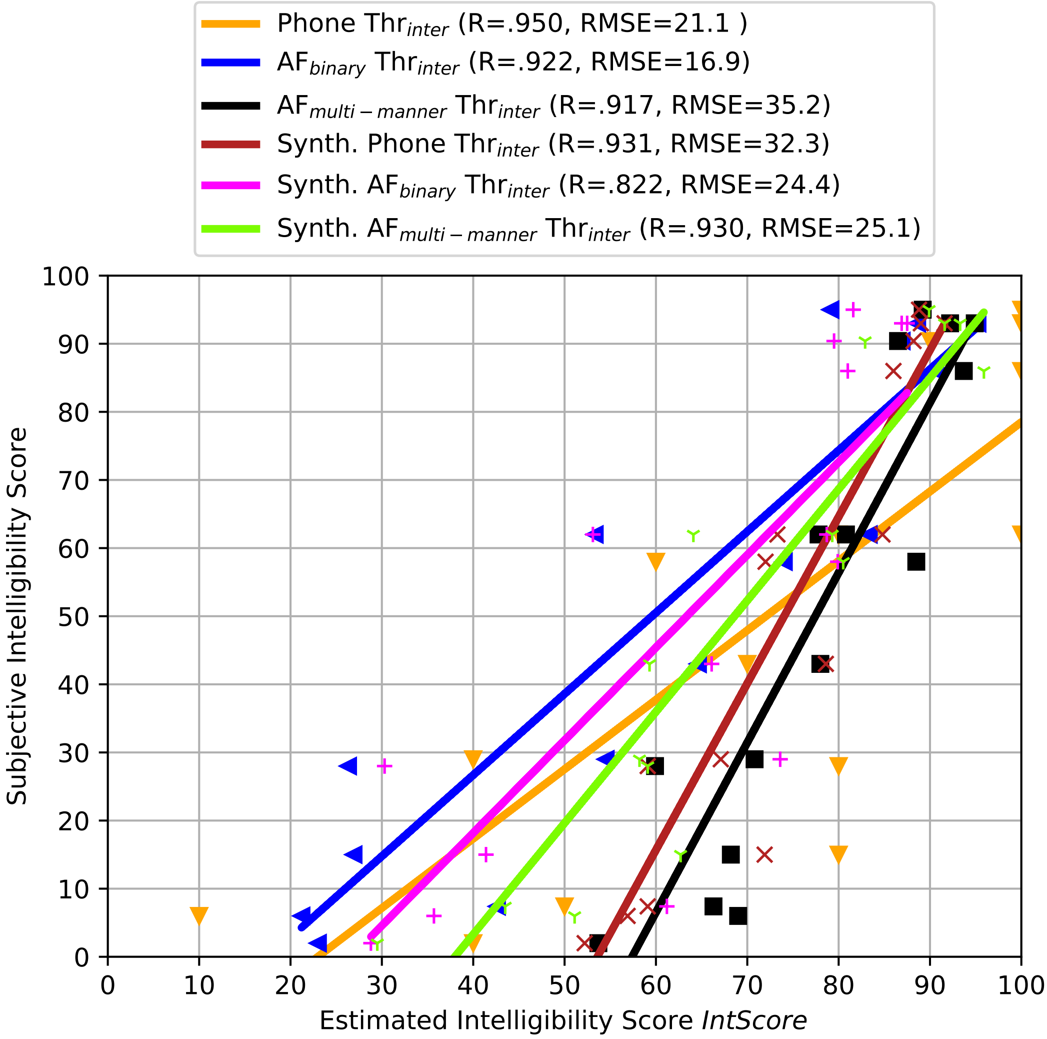


Figure 5.4: Pearson's correlation plot obtained from proposed intelligibility assessment systems. Synth refers to the case of a single synthetic control.

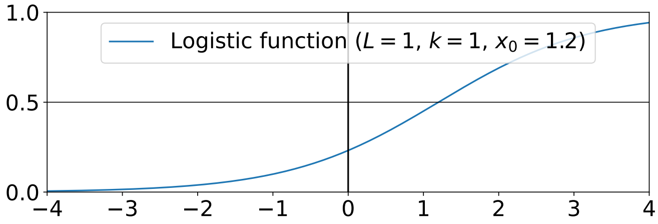


Figure 5.5: Logistic function used to convert a DTW match score $L(M,N)$ into a posterior probability.

proposed probabilistic atypical utterance verification results. The results are consistent with the previous results, where the phone space and AF$_{multi-manner}$ space are performing better than the AF$_{binary}$ space. The results also demonstrate that with the proposed probabilistic approach it might be possible to reduce the necessary amount of in-domain recordings, given an estimate of the same-word distribution's midpoint is available, e.g. from a different experiment with the same representation.

Table 5.4: Pearson's correlation ($r$) and Spearman's correlation ($\rho$) between subjective and objective intelligibility from posterior estimates. $p$-values are presented in Italics font.

| Posterior feature space | $Thr_{P_c}$ | |
|---|---|---|
| | $r$ | $\rho$ |
| Phone | .854 | **.965** |
| | *4.90e-5* | *6.11e-9* |
| AF$_{binary}$ | .550 | .594 |
| | *3.33e-2* | *1.94e-2* |
| AF$_{multi-manner}$ | .781 | .781 |
| | *5.73e-4* | *5.75e-4* |

### 5.3.5 Non-posterior representations for utterance verification

While the previous sections have demonstrated the effectiveness of the proposed approach with posterior representations, in the paralinguistic field, many representations are used, which we propose to investigate briefly. Mel frequency cepstral coefficients (MFCC) (Hönig et al., 2005) are one of the most common representations for speech analysis. Additionally, we used the wav2vec 2.0 representation (Baevski et al., 2020), an embedding representation developed for speech recognition and trained in a self-supervised manner on large amounts of unlabeled speech data. We used the wav2vec 2.0 base model[1], which is a 768-dimensional representation. Table 5.5 shows the results of estimating intelligibility with the aforementioned representations. Since the representations are no longer probability distributions, we change the distance metric. MFCCs perform well, when multiple healthy references are available, but fail, once synthetic speech is used as a reference. Unexpectedly, wav2vec 2.0 yields a bad performance with euclidean and cosine distances.

## 5.4 Analysis of acoustic differences after lip filler surgery

In this section, we validate the proposed approach on a different atypical speech aspect, where systematic changes occur in a particular part of speech production system. In collaboration with Univ.-Prof. DDr. Kurt Alexander Schicho from the Medical University of Vienna (MUV)[2], we investigated the effects of cosmetic lip filler surgery on speech production. This pilot study aimed to propose a method to determine whether cosmetic lip filler surgery is altering speech production capabilities, e.g. due to loss of sensation in the lips and the effect on producing plosives. To that end, we contrast a subjective listening test with the proposed objective

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/README.md

[2] https://schicho-medical.at

Table 5.5: Pearson's correlation ($r$) and Spearman's correlation ($\rho$) between subjective and objective intelligibility estimates for non-posterior representations. DM denotes the distance metric. $p$-values are presented in Italics font.

| Representation | $Thr_{inter}$ | | DM |
|---|---|---|---|
| | $r$ | $\rho$ | |
| MFCC | .923 | .940 | euclidean |
| | *9.23e-07* | *1.75e-07* | |
| Synthetic ref. MFCC | .514 | .669 | euclidean |
| | *4.99e-2* | *6.37e-3* | |
| wav2vec 2.0 | -.558 | -.501 | euclidean |
| | *3.05e-1* | *5.71e-2* | |
| wav2vec 2.0 | -.290 | -.175 | cosine |
| | *2.935e-1* | *5.31e-1* | |

intelligibility assessment: i) an A-B listening aiming to see, if German native listeners can distinguish between before and after surgery (cf. Section 5.4.2) and ii) an automatic method: the average distances of recordings of the same speaker before and after surgery compared to the distance of recordings of different speakers (cf. Section 5.4.3).

### 5.4.1 Data

For our studies, MUV provided recordings from 11 women before and after having undergone cosmetic lip filler surgery. All 11 subjects read the following 5 minimal pairs:

Table 5.6: Minimal pairs for MUV data collection.

| | | |
|---|---|---|
| 1 | Baum | Traum |
| 2 | Vater | Pater |
| 3 | Nüsse | Küsse |
| 4 | See | Tee |
| 5 | Lampe | Tante |

Additionally, all subjects read the following text passage from Hermann Kant's novel Lebenslauf, zweiter Absatz: Erzählungen" (Kant, 2011). The recrodings were segmented into 7 phrases:

### 5.4.2 A-B listening test

To identify any difference in pronunciation capabilities are perceivable, an A-B preference test was conducted. Only Austrian and southern German native speakers were selected as listeners. A total of 20 listeners participated. Listeners were asked to pick a preference in terms of clarity of pronunciation. Even though the instruction was to choose a preference, some participants opted not to pick any, which, in the analysis, will be considered as "no preference". The test started with an example of minimal pairs and sentence reading, where the listeners know A

Table 5.7: Text passage for MUV data collection from Hermann Kant's novel Lebenslauf, zweiter Absatz: Erzählungen".

1 Ich saß auf dem Dach und konnte alles genau sehen:
2 die vier verstaubten Männer in der Buchenlaube,
3 meine Mutter und die Frau mit der Ziege,
4 meine kleine Schwester Alida hinter dem Schattenmorellenspalier,
5 den Festzug mit Blumen und Fahnen in der kleinen sandigen Sprache und Judith, die Königin.
6 Die Königin stand ganz alleine auf dem sauber geharkten Weg zwischen dem Steingarten und der Dahlienreihe.
7 Sie wartete auf den König.

is the stimulus before, and B is the one after "some form of treatment". Following, 33 tests are provided, two read sentences and one minimal pair that were randomly selected from all recordings of all 11 subjects. Every time the test is opened, both the order of the tests was randomized, and so were before and after behind the A and B stimuli. For the evaluation, we confirmed the conscientiousness of listeners, meaning that we would exclude listeners that picked only A or B.

**Overall**, a Wilcoxon signed-rank test for the null hypothesis[3], we observed a p-value of $p = .6524$, whereas a Student's t-test[4] gives a p-value of $p = .5141$, when the expected population mean is set to the number listeners divided by two ($popmean = 10$). Hence, it can be concluded that there is no observable difference between before and after the lip filler surgery.

**Evaluation per speaker**: To test, whether listeners could discriminate significantly between before and after on a per-speaker-level, we evaluated significance tests per speaker: To all 20 listeners For every subject of the study, 3 tests were given, so 60 tests per speaker. Table 5.8 shows p-values that were obtained with a Student's t-test, where the expected population mean is again the number of listeners divided by two ($popmean = 10$).

Table 5.8: Table 1: Student's t-test of the preference test per speaker; per speaker, 3 samples were given to 20 listeners.

| Speaker | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| p-value | .6256 | .9096 | .9096 | .5 | .3440 | .5 | .8304 | .8304 | .2048 | .7422 | .2048 |

Finally, in Figure 5.6, we visualized the listeners' preferences in a bar plot. As for the hypothesis testing, preferences were aggregated over 3 tests per 20 listeners, in total 60 tests per speaker.

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html

[4] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_1samp.html

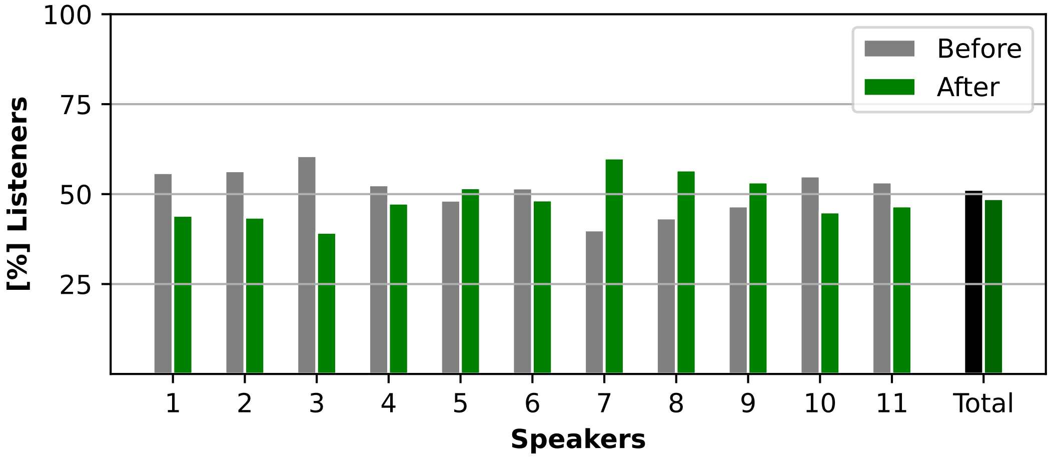


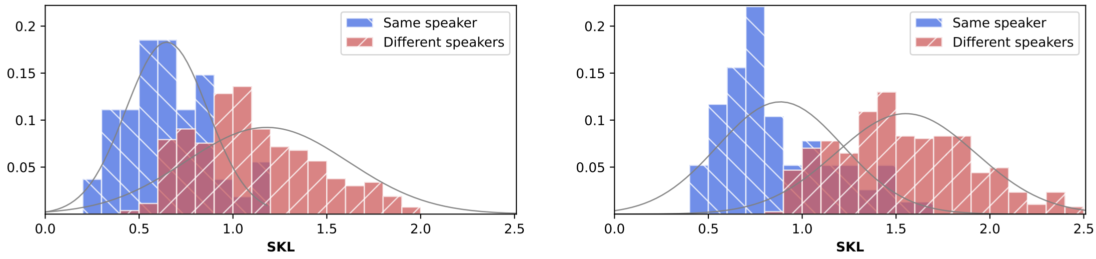
Figure 5.6: Percentage of preference in an A-B listening test per speaker and in total.

### 5.4.3 Utterance verification-based assessment

In addition to the listening test, we proposed an automatic analysis method: Following the approach described in Section 5.2, we proposed to assess whether a subject's speech after surgery deteriorates from before's by matching considering the utterance before as reference and matching the utterance after the surgery to obtain a distance. As a point of reference, we match the all same utterances from before the surgery to different speakers' recordings of the same phrase before the surgery. The hypothesis is, that if the distance between the before and after of the same speaker is lower than the distances to different speakers, i.e. inter-speaker-variability, we can assume that the surgery went well. We acknowledge that comparing matching scores of entire sentences is a crude approach. Therefore, in the following, the results on the minimal pairs as our main indicator, but report the results on the sentences for the sake of completeness.

The experiment is carried out in phone posterior space (cf. Section 5.3.1) and the SKL distance metric is applied since we observed consistent performance with it. Figure 5.7 shows histograms of all distances between same-speakers before to after as all as distances of different speakers of the same word. Over the histogram, a normal distribution was overlaid to better visualize the mean and standard deviation of the two sets of distances. Figure 5.7 (a) shows these same histograms for individual minimal pair recordings, Figure 5.7 (b) for read sentences. In both figures, same-speaker distributions have a lower mean than different-speaker distributions, which confirms the results of the listening test, that after the surgery no acoustic differences are perceivable.

Analogous to the listening test analysis, we also tested whether the difference between same-speaker and different-speaker distances differ on a speaker level. Figure 5.8 shows the average same-speaker- and different-speaker distances per speaker. We observed that consistently the same-speaker distances are lower than the different-speaker distances. From this, we again

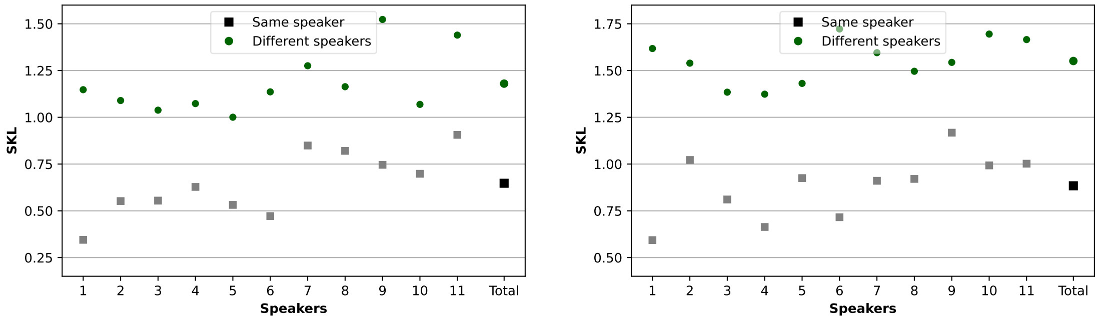


(a) Minimal pairs

(b) Sentences

Figure 5.7: Histogram of same-speaker- and different-speaker SKL-distances.

conclude no noticeable acoustic differences between before and after the surgery.



(a) Minimal pairs

(b) Sentences

Figure 5.8: Averages of same-speaker- and different-speaker SKL-distances.

## 5.5 Summary

In this chapter, we presented a novel approach for atypical intelligibility assessment. Framing intelligibility assessment as a divergence from healthy speech and solving it through utterance verification is novel and intuitive. Both individual matching score distances and aggregated intelligibility scores are explainable to clinicians, which is an advantage over other existing approaches.

The approach was evaluated on two small but typical-size datasets. On the UA-Speech dataset, we achieved competitive results, while not requiring any training, and by using auxiliary posterior estimators. Notably, our approach is not limited to the used representation, meaning that with an improved representation (and the right distance measure) it can be improved in the future. Additionally, we showed, how simple and robust the method is to reducing the amount of control recordings. As a consequence, the same idea could be applied to an open research question: For cosmetic lip filler surgery, we verified the results of an A-B listening test. Implicitly, in both cases, we close the production-perception loop by replacing a listening test with examples of control speech.

# 6 Pronunciation assessment of atypical speech

This chapter extends Chapter 5 by proposing a method for pronunciation assessment. We again rely on the notion that atypical speech can be measured as a deviation from healthy control speech. We propose a method, that again compares atypical utterances to a reference sequence, which allows us to determine which sub-word units, i.e. phonemes, are pronounced correctly and which are not. A sub-word unit pronunciation assessment is an upgrade of the need of end-users. Speech therapists and patients want to identify speaking deficits to improve specific speech production aspects.

## 6.1 Related works

Different methods have been proposed for pronunciation assessment. In the field, second language (L2) learning and assessment for people with speaking disorders are the two main use cases that are considered. In the following, we briefly overview popular methods.

Traditionally, goodness of pronunciation (GOP) has been used for evaluating pronunciation assessment (Witt and Young, 2000). Based on a phoneme alignment, a GOP value can be computed for every phoneme. GOP is defined as the duration-normalized posterior probability ratio (sometimes also likelihood ratio) between the realized phone and the expected/canonical phone:

$$GOP(p) \approx log \frac{P(p|o; t_s, t_e)}{max_{q \in Q} P(o|q; t_s, t_e)} / NF(p) \tag{6.1}$$

Given an expected phoneme $p$, the observation $o$, $t_s$ and $t_e$ are the start and end time of the phoneme and $NF(p)$ the length of phone $p$. $P(o|q); t_s, t_e$ are the likelihoods of phone $q$. $Q$ is the set of all possible phones. The maximum likelihood is computed over all possible phones. The GOP value is then used to determine whether a phoneme is pronounced correctly or not. To make an assessment, a phone-based threshold is proposed as $Thr_p = \mu_p + \sigma_p$. Another generic option of setting a threshold is based on an equal error rate, as done in (Wei et al., 2022).

GOP, in the context of L2 learning, is also adopted with end-to-end ASR systems, for example by Zhang et al. (2020) with a CTC-based ASR system, which outputs character sequences and GOP is computed based on that. In (Korzekwa et al., 2021), the authors propose a system for second language learners. For word-level correctness assessment, the authors propose to train a neural network in an end-to-end manner. For phoneme-level feedback, the authors propose to append a phoneme-recognizer that outputs the predicted sequence.

A variety of methods on pathological intelligibility estimation is proposed in (Middag, 2012): Based on an ASR-alignment, phonological features as input to different regression methods are proposed (e.g. linear regression and support vector regression) to estimate intelligibility for all speakers in a dataset. Requiring to train on a certain number of speakers is a drawback of this method. Additionally, even though speech production variables like articulation and phonation quality are predicted, no phoneme-level assessment is attempted. In (Quintas et al., 2022), the authors are estimating the intelligibility of head and neck cancer patients by force-aligning speech, and separately training one Siamese network on similarity estimation for 16 different consonants. Intelligibility is estimated based on the similarity scores.

## 6.2 Proposed Kullback-Leibler divergence-based Hidden Markov Models

In order to model healthy speech, we propose to use a Hidden Markov Model (HMM) with a Kullback-Leibler divergence cost function (KL-HMM). We propose to train HMMs per phoneme on control recordings, which will serve as a reference. In the following, we elucidate the KL-HMM training:

In the context of hybrid HMM/ANN speech recognition, HMMs are used for acoustic sequence modeling, while neural networks ANN are used to predict posterior probability distributions, e.g. over context-independent phones, typically denoted as $a$, belonging to an observation $x$: $p(a|x)$. Traditionally, HMMs were parametrized with Gaussian mixture models that emit the likelihoods of observations. In Aradilla et al. (2007) a modification was proposed: a posterior-based HMM model, where the target parametrization is categorical probability distributions. As observations, posterior probability distributions are used. The HMM's cost function is a KL divergence cost function between observation and categorical distributions:

$$KL(y||z) = \sum_{k=1}^{K} y_k \log \frac{y_k}{z_k} \tag{6.2}$$

where $y$ and $z$ are two probability distributions, and $K$ is the number of states in the HMM. The KL divergence is a measure of the difference between two probability distributions. To train an HMM, KL-divergence is minimized:

$$J_\theta = \min_\theta \sum_{t=1}^{T} KL(y_{\theta_t} || z_t) \tag{6.3}$$

The parameters of the KL-HMM are estimated with the Viterbi EM algorithm by minimizing $J_\theta$: For training, i) data is segmented uniformly according to the phonetic transcriptions, ii) target distributions are computed to minimize the KL-divergence according to the segmentation, iii) training data is segmented to minimize the global cost function using the Viterbi algorithm. Steps ii) and iii) are repeated until convergence of the cost function.

In practice, the following resource is required: Training utterances with phonetic transcriptions according to a lexicon; we use the phonemes as phonetic units. Per phonetic unit, we train a 3-state HMM; each state's target posterior distribution gets updated, and states are not tied to particular observation but are only characterized by their target posterior. We're not making use of the HMM's transition probabilities.

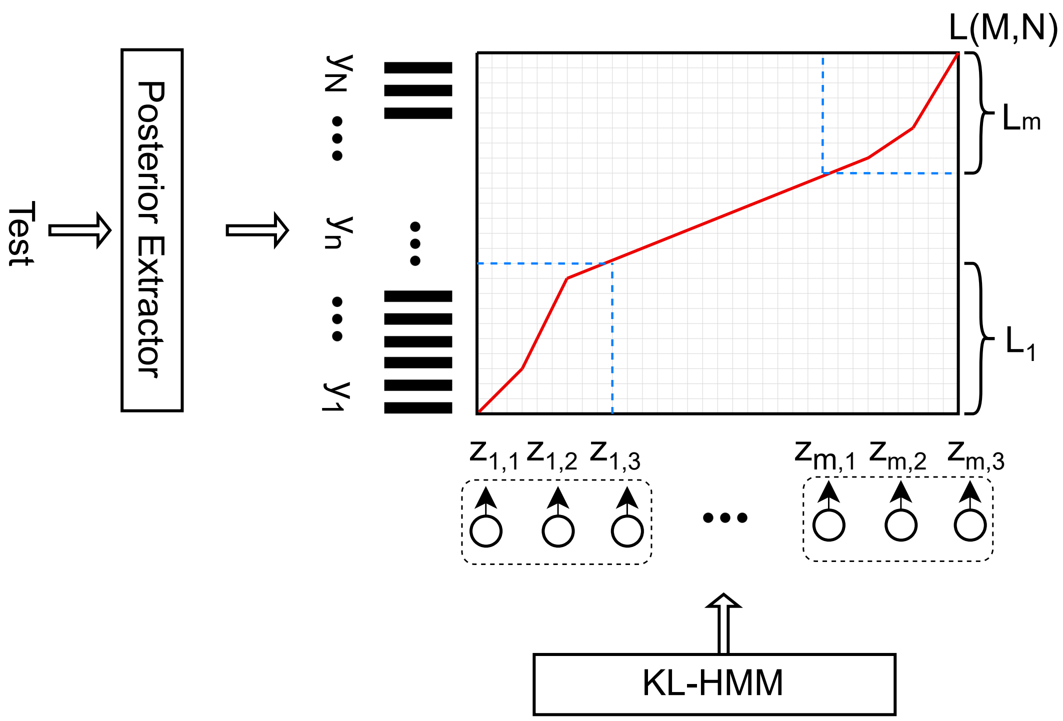


Figure 6.1: Alignment to KL-HMM's phonetic units through DTW path.

When a word's phonetic unit's Kl-HMM-target posterior distributions (per phonetic transcription) are concatenated together, we obtain a test utterance's expected sequence. Matching the two is illustrated in Figure 6.1. For all test frames belonging to one KL-HMM phonetic unit, the local distances can be accumulated and path-length-normalized to obtain a distance between that phonetic unit and frames in the test utterance (cf. $L_1$ and $L_m$ in Figure 6.1). Thus, we get an alignment for the test utterance as well as a distance between the KL-HMM reference phoneme sequence and corresponding frames of the test utterance. These phoneme-level scores can then be used for 'phoneme-verification' – in contrast to the previous utterance verification. On a side note: The individual phoneme-level distances sum up to the total match score $L(M,N)$.

## 6.3 Experimental study on UA-Speech

As an initial experiment and extension of the results in Chapter 5, we trained a KL-HMM on the UA-Speech database (cf. Section 5.3.2). We perform this experiment in phone posterior space (cf. Section 5.3.1) since it yielded the best results in Chapter 5. Note, that we used both male and female speakers for training the HMM models. The intersection-point threshold is obtained by creating same-word and different-word pairs between KL-HMM sequences and control speech. For testing, only one reference is available: the KL-HMM posterior sequence. Same as in Section 5.2, intelligibility is estimated as the number of words that are considered pronounced correctly. The performance presented in Table 6.1 is comparable to the results in Chapter 5. The result demonstrates, that we can model control speech with a KL-HMM. However, on UA-Speech, we can only evaluate performance globally, since the individual utterances were not annotated in terms of how pronunciation was realized.

Table 6.1: Pearson's correlation ($r$) and Spearman's correlation ($\rho$) between subjective and objective intelligibility estimates on UA-Speech with a KL-HMM reference model. $p$-values are presented in Italics font.

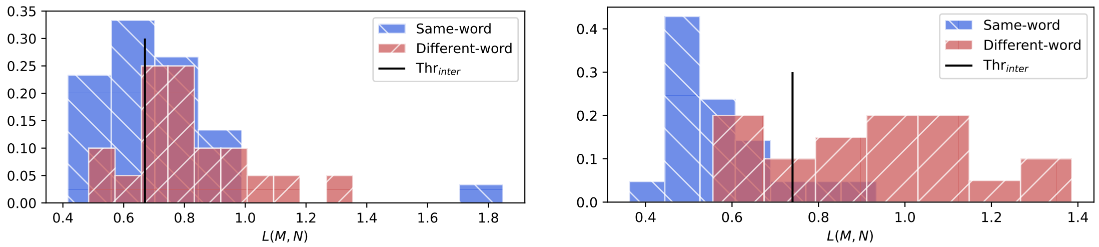
| | $Thr_{inter}$ | |
|---|---|---|
| | $r$ | $\rho$ |
| Phone KL-HMM | .936 | .910 |
| | *1.79e-7* | *2.42e-6* |

## 6.4 Experimental study on Torgo

We conducted a set of experiments on the Torgo dataset. Torgo has recordings of speakers with dysarthria and phonetic transcription for most utterances, which allows a word-level and phoneme-level assessment. The main goals of this experimental study on Torgo are (i) utterance verification-based intelligibility assessment, and (ii) performance evaluation of phoneme-level pronunciation assessment. Both can be performed with either a data-driven threshold $Thr_{inter}$ or a non-data-drive threshold by converting distances into posterior probabilities and using $Thr_{P_c} = .5$. For (i) we distinguish the following decision levels:

**Global:** One global threshold for utterance-verification-based intelligibility estimation (cf. Section 5.2) on Torgo. We use the experimental same setup as in Chapter 5, but with the Torgo database.

**Word:** In order to improve the utterance-verification-based intelligibility estimation setup proposed in Chapter 5, we propose an initial variation: Instead of using a global utterance verification threshold, we propose to use one for every unique word in the database. This seems reasonable because therapists let their patients practice a fixed set of words. The word-specific thresholds are obtained by creating same-word and different-word pairs from control speech. In Torgo, that means creating word pairs for 600 individual words. Figure 6.2 shows an example of the histograms for two similar words, where both distributions and thresholds differ slightly, which lets us expect performance gains.



(a) Word: feet

(b) Word: feed

Figure 6.2: Histogram of same and different-word pair scores $L(M, N)$ for two words in the Torgo database.

**Phoneme:** For pronunciation assessment on the Torgo dataset, we train KL-HMMs on Torgo's control utterances. Based on the obtained alignment (cf. Figure 6.1) between KL-HMM and test sequence obtained through DTW and the phoneme-level thresholds. Same-phoneme distances are obtained by matching control utterances to the corresponding KL-HMM sequence. Different-phoneme distances can be obtained for every phoneme, by matching the KL-HMM test sequence to words that only differ in that phone, hence having a Levenshtein distance of 1 in terms of phonetic transcription.

### 6.4.1 Dataset

**Torgo:** The Torgo corpus of dysarthric speech (Rudzicz et al., 2012) contains English recordings from 15 speakers, of which 8 are severely dysarthric speakers and 7 control speakers. For our experiments, we included only the isolated word recordings. Of all isolated words, from control speakers, we further exclude those without a transcription; from the dysarhtric group, we only include recordings for which a perceived phonetic transcription is available, since it is needed for evaluation. A word is considered pronounced correctly when the target and perceived pronunciation are the same. That leaves a total of 600 unique words. Table 6.2 overviews the number of utterances for control and dysarthric speakers. The speakers with dysarthria have cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS), covering a wide range of intelligibility levels. We used cmu-dict pronunciation dictionary for our experiments.

Table 6.2: Torgo corpus statistics of utterances that are isolated words and that have a phonetic transcription.

| | # Utterances | Length |
|---|---|---|
| Control | 4359 | 2h32min |
| Dysarthric | 897 | 30 min |

### 6.4.2 Results

**Utterance verification-based intelligibility assessment:** Table 6.3 summarizes the results on Torgo of word-level intelligibility assessment at different decision-levels. We are not reporting correlations, since only 7 speakers have contributed isolated words and correlations wouldn't be statistically significant. We instead report root mean squared error (RMSE) between the estimate and annotated global intelligibility, as well as word accuracy (WAC) as the percentage of utterances that are identified correctly as mispronounced/correctly pronounced. Overall, the probability threshold $Thr_{P_c}$ consistently outperforms $Thr_{inter}$. Contrary to our expectation, we do not observe performance gain when performing utterance verification with a word-specific threshold. With a phoneme-specific threshold, where a word is considered wrong when one of its phonemes is mispronounced, we observe slightly more robust results. Yet, a phoneme-specific threshold is less computationally heavy than one per word and offers other advantages.

Table 6.3: RMSE and word accuracy (WAC) between subjective and estimated intelligibility on Torgo.

| Reference | Threshold | Decision-level | RMSE | WAC [%] |
|---|---|---|---|---|
| Recordings | $Thr_{inter}$ | global | 49.24 | 45.30 |
| Recordings | $Thr_{P_c}$ | global | **19.54** | 74.91 |
| Recordings | $Thr_{inter}$ | word | 53.60 | 40.07 |
| Recordings | $Thr_{P_c}$ | word | 21.47 | 77.30 |
| KL-HMM | $Thr_{inter}$ | phoneme | 19.78 | 62.20 |
| KL-HMM | $Thr_{P_c}$ | phoneme | 22.09 | **77.64** |

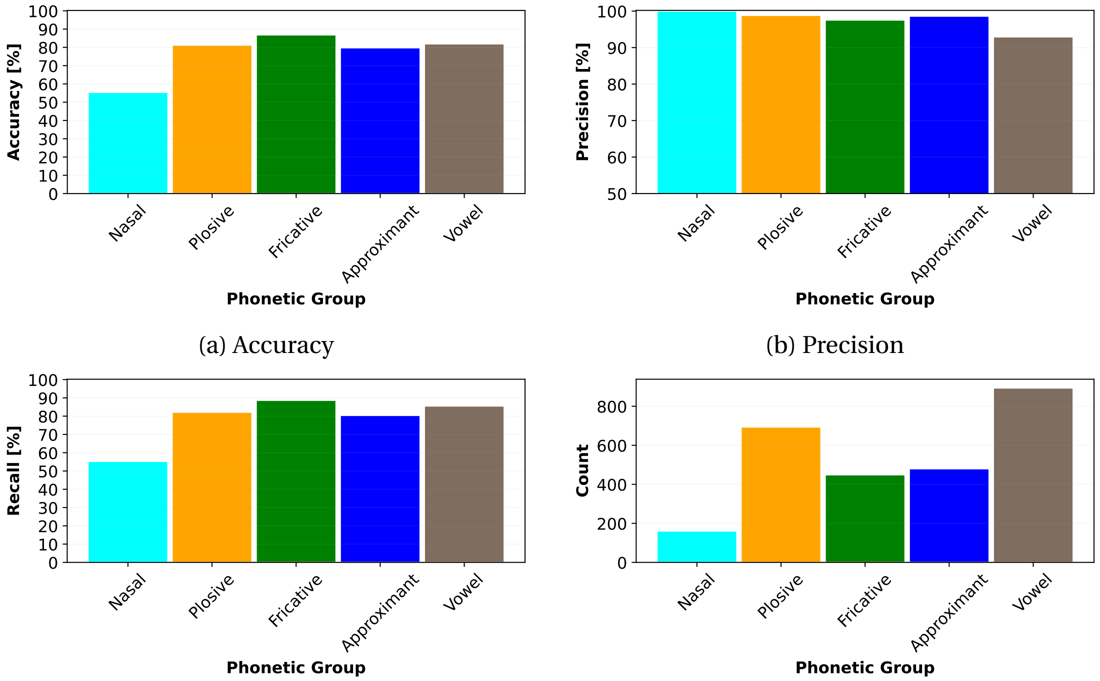
**Phoneme-level pronunciation assessment:** In order to measure the performance of the phoneme-level pronunciation assessment, we directly assess performance on a phoneme level. We use the posterior-based threshold $Thr_{P_c} = .5$. We calculate phoneme accuracy, precision, and recall between our target and perceived pronunciation and our decisions. Table 6.4 shows overall performance. Over all speakers and all phonemes, we achieve a 77% accuracy.

Table 6.4: Accuracy, precision and recall of phoneme-level pronunciation assessment in Torgo.

| Accuracy [%] | Precision [%] | Recall [%] | Count |
|---|---|---|---|
| 77.03 | 97.59 | 78.42 | 2680 |

### 6.4.3 Analysis

Based on the results in Table 6.4, we split the phonemes into different groups and analyze the performance of the system in different groups. In Figure 6.3, we show accuracy, precision, recall, and count for the phonetic group's nasals, plosives, fricatives, approximants, and vowels. On accuracy and recall, we observe lower performance of 55% for nasals but accuracies above 80% for the remaining phonetic groups.



(a) Accuracy

(b) Precision

(c) Recall

(d) Count

Figure 6.3: Phoneme-level performance for different phonetic groups in Torgo, as well as total count of phonemes used in the analysis.

## 6.5 Experimental study on COPAS

In this experimental study, we use the Dutch COPAS dataset, because it is a large dataset with a variety of pathological speech and thereby gives a stronger supporting argument to our proposed approach. We follow the same experimental design as in Section 6.4, meaning we are interested in (i) utterance verification-based intelligibility assessment, and (ii) performance evaluation of phoneme-level pronunciation assessment. We apply either a data-driven threshold $Thr_{inter}$ or a non-data-driven threshold by converting distances into posterior probabilities and using $Thr_{P_c} = .5$. Again, for (i) we distinguish the following decision levels: **global**, **word**, and **phoneme**.

### 6.5.1 Dataset

**COPAS:** Dutch corpus of pathological and normal speech (COPAS) is a database of speech recordings published by Van Nuffelen et al. (2009). For our experiments, we use the isolated word recordings called Dutch Intelligibility Assessing (DIA) which is a designed list of consonant-vowel-consonant words. We consider words, for which a perceived pronunciation was annotated. A word is considered pronounced correctly when its target pronunciation and perceived pronunciation are the same. In total, COPAS has 319 speakers of different groups of which we include: healthy speech as references and 7 impaired groups: voice disorder, cleft, articulation disorders, laryngectomy, glossectomy, and dysarthria. Our final word list which is both in train and test contains 715 different words. The pronunciation dictionary was provided by the publishers of the dataset.

Table 6.5: COPAS corpus statistics of utterances that are isolated words and that have a phonetic transcription.

| Group | # Speakers | # Utterances | Avg. Intelligibility |
|---|---|---|---|
| Voice Disorder | 8 | 400 | 88 |
| Larynx | 37 | 1846 | 75 |
| Dysarthria | 87 | 4349 | 80 |
| Cleft | 38 | 1849 | 86 |
| Articulation Disorder | 3 | 150 | 94 |
| $\sum$ | 173 | 8594 | 84.6 |
| Control | 78 | 5836 | 94 |

### 6.5.2 Results

Since we evaluate our approach on 173 speakers, we measure performance in terms of Pearson's correlation r, Spearman's correlation $\rho$, RMSE between true and estimated global intelligibility, and WAC as the percentage of utterances that are identified correctly as mispronounced/correctly pronounced. As a baseline, we consider results from Middag (2012), who trained a linear regressor from phonological features on the entire speaker set of the COPAS DIA set in a cross-validation evaluation. The authors achieve a Pearson's correlation of $r = .814$ and an RMSE of 7.71 (cf. Table 8.1 in Middag (2012)).

**Utterance verification-based intelligibility assessment:** The experiments presented in Table 6.6 are conducted in phone posterior space. We choose it because it has proven to perform well and neglect the language mismatch - COPAS is dutch - our phone posterior representation is trained on English data. Our overall best results achieved are a Pearson's correlation of r= .357, RMSE of 17.68 and a WAC of 77.54%. We achieve lower WAC with KL-HMM references, presumably because different pronunciation variations can be compensated better with multiple human references that are more indicative than one single 'average' reference. The baseline performance far exceeds our result. However, their method trains a regressor instead of our method of counting and only allows a global estimate, and no utterance-level feedback to give to speakers.

Table 6.6: Pearson's correlation r, Spearman's correlation $\rho$, word accuracy (WAC), and RMSE between subjective and estimated intelligibility on COPAS.

| Reference | Threshold | Decision-level | r | $\rho$ | RMSE | WAC [%] |
|---|---|---|---|---|---|---|
| Recordings | $Thr_{inter}$ | global | .2991 | .1993 | 18.72 | 69.15 |
| Recordings | $Thr_{P_c}$ | global | .3008 | .2282 | 35.82 | 53.95 |
| Recordings | $Thr_{inter}$ | word | .314 | .2311 | 22.97 | 63.28 |
| Recordings | $Thr_{P_c}$ | word | .351 | .289 | **17.68** | **77.54** |
| KL-HMM | $Thr_{inter}$ | phoneme | **.357** | **.294** | 26.34 | 57.28 |
| KL-HMM | $Thr_{P_c}$ | phoneme | -.058 | -.001 | 79.09 | 20.75 |

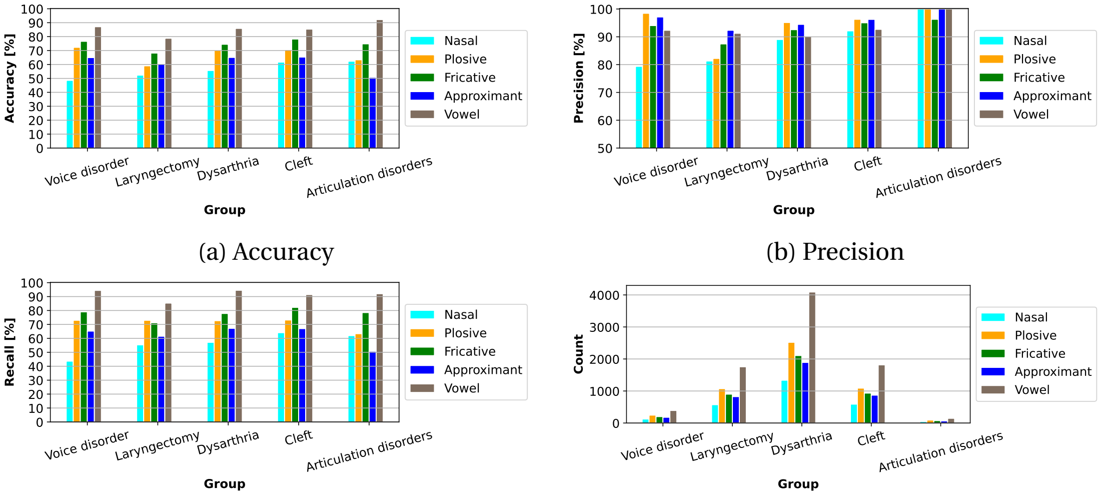
**Phoneme-level pronunciation assessment:** On top of the global intelligibility, we present our results on the performance of phoneme-level pronunciation assessment. We used the intersection-point threshold $Thr_{inter}$. On average, we observe a performance drop of 5-10% when compared to the results on Torgo. Typically, a larger number of samples has a larger variance, hence is more difficult. Per-group, the lowest phoneme accuracy is achieved for the laryngectomy group, which also has the lowest average intelligibility of 75, whereas the groups with intelligibility above 80 achieve accuracies around 70% (cf. Table 6.5). This could mean increased difficulty in stronger severity disabilities.

Table 6.7: Accuracy, precision, and recall of phoneme-level pronunciation assessment in COPAS.

| | Accuracy [%] | Precision [%] | Recall [%] | Count |
|---|---|---|---|---|
| Voice disorder | 69.92 | 92.27 | 71.07 | 1145 |
| Laryngectomy | 63.74 | 86.91 | 69.27 | 5130 |
| Dysarthria | 70.27 | 92.28 | 73.89 | 11951 |
| Cleft | 72.26 | 94.48 | 75.59 | 5303 |
| Articulation disorders | 68.71 | 99.25 | 69.34 | 430 |
| $Avg.$ | 68.98 | 93.04 | 71.83 | $\sum = 20949$ |

### 6.5.3 Analysis

Finally, in Figure 6.4, we illustrate performances for every group but differentiate between phonetic groups. We use the result achieved with $Thr_{inter}$. We observe, consistently over the disorder groups, nasals have the lowest accuracy between 50 %and 60%, followed by approximants. Plosives and fricatives are recognized with an accuracy of around 70%, while for vowels it ranges 80% to 90%. The same trends can be seen in precision and recall. One potential reason for the lower performance in nasals could be that, as discussed in Section 2.1, nasality is one of the four main intelligibility dimensions, and a representation more tuned to nasality is necessary.



(a) Accuracy

(b) Precision

(c) Recall

(d) Count

Figure 6.4: Phoneme-level performance for different phonetic groups of the groups of speakers in COPAS, as well as total count of phonemes used in the analysis per group.

## 6.6 Summary

In this chapter, we extended our novel utterance verification approach for intelligibility assessment by varying the decision level and observe good results when it is word-specific. We introduced a new way to model control speech with phoneme KL-HMM models. Again, we implicitly approached pronunciation assessment as a divergence from healthy speech, yet now on phoneme-level. We demonstrated the effectiveness of this method on two datasets: Torgo, a set of recordings from English-speaking people with dysarthria, as well as COPAS, a set of recordings from dutch-speaking people with different speech production disabilities. By demonstrating good performance on both datasets, we confirmed the robustness of the approach across languages while still using auxiliary resources for estimating posterior representation.

# 7 Conclusion and future directions

Analysis of atypical aspects in speech is a challenging and very interesting research topic. It includes social and medical knowledge in the analysis of human speech production. The goal of the thesis was to propose novel methods for detection and analysis that are tailored to the problem just as much as to the needs of the end users. We focused on guiding the learning of CNN-based architectures, to improve performance as well as to better cope with the low-resource aspect that is common in atypical speech assessment. For intelligibility estimation, we introduced a new scheme that is interpretable. The method is extendable to phoneme-level pronunciation feedback and should therefore be appealing to both clinicians and patients.

In Chapter 3, we investigated inducing knowledge into CNNs for classifying and measuring the severity of Alzheimer's disease. We found a shorter first-layer kernel width to be more suitable, as well as using a filter, the ZFF, to pre-filter signals before feeding them to the networks to be more efficient than using the unfiltered raw waveform. In research collaborations, we discovered that a late fusion of acoustic and text-based systems yields performance gains, indicating that they model different information. For the degree of sleepiness estimation, we applied the CNN framework as well, where the initial performances were sub-par. However, a pre-training scheme to predict articulatory features helped. Finally, when combining different models, we achieved a robust estimate.

In Chapter 4 we investigated phonation assessment in a breathing impairment task of Parkinson's patients. We compared hand-crafted features that model voice-source information, as well as feeding the signals into raw waveform CNN models. Input to all systems were different filtered signals. We found that CNNs learn good embedding representations. Functionals of these embeddings outperform hand-crafted features. Furthermore, we found that breathing impairment may not need to be analyzed in glottal signals, since raw and ZFF signals yielded the best systems.

In Chapter 5 we measured intelligibility, a popular clinical measure of the severity of pathological speech. It is an aggregate over a set of utterances of a speaker, and we, therefore, proposed to deal with it as such. We implicitly emulate human listeners' assessments. We proposed verifying the correctness of the pronunciation of every utterance and aggregating those decisions into the intelligibility score. The verification of utterances is performed in different auxiliary-resource acoustic-phonetic posterior spaces. Overall we found that it compares well to other methods in the literature. Several variations, e.g. in terms of the amount of control speech demonstrated the approach's robustness. Additionally, we were able to demonstrate, that the same idea can be applied to confirm the results of an A-B-listening test on recordings of before and after lip-filler surgery.

In Chapter 6, we extended the idea of intelligibility estimation to pronunciation assessment. Instead of verifying words, we propose to verify phonemes, and can therefore give pronunciation feedback with high accuracy. We supported our proposed approach with experiments on two datasets. We discovered the accuracy varies across phonetic groups. Specifically, nasals are assessed less accurately.

Regarding the general limitations of the proposed approaches, we acknowledge that even though the objective nature of automatic methods is an advantage, they will in the near future rather serve as a second opinion. Pronunciation assessment can be used for patients wanting to exercise at home. Still, these patients have to be supervised by trained therapists.

In the following, we propose possible directions for future research:

- The detection of pathologies in speech so far focused on directly classifying from utterance-level representations. In our opinion, analysis can be improved by taking a more holistic approach: A combination of speech dimensions, such as articulation, nasality, phonation, and prosody. This could lead to better objective assessments that are also more interpretable.
- The proposed methods for intelligibility estimation and pronunciation assessment may benefit from better auxiliary speech representations, e.g. self supervised learning-based representations. In this work, we investigated representations that were not fine-tuned in a task or domain specific manner (e.g., phonetic classification). So, a question that emerges for future research is: to what extent the proposed methods can benefit from task or domain specific adaptation?
- We believe that the presented method for pronunciation assessment is suitable for speech therapy, but future work is needed to bring performance up to an applicable level. This could also be done through more threshold tuning, e.g. per group, or even personalized.

# A Phoneme-to-articulation mapping

Table A.1 shows a phoneme-to-articulatory feature map. It is mapping all 54 English phonemes, in ARPABET notation (Klautau, 2001), to 5 articulatory categories. The categories are manner, place, height, and vowel. The mapping is based on Rasipuram and Magimai.-Doss (2016). We note, that such a mapping is not unique.

Table A.1: Knowledge-based phoneme-to-articulatory feature map used in Section 3.3.5.

| Phoneme | Manner | Place | Height | Vowel |
|---|---|---|---|---|
| sil | sil | sil | sil | sil |
| aa | vowel | back | low | aa |
| ae | vowel | mid-front | low | ae |
| ah | vowel | mid | mid | ah |
| ao | vowel | back | mid-low | ao |
| aw1 | vowel | mid-front | low | aw1 |
| aw2 | vowel | mid-back | high | aw2 |
| ax | vowel | mid | mid | ax |
| axr | approximant | retroflex | mid | consonant |
| ay1 | vowel | back | low | ay1 |
| ay2 | vowel | mid-front | high | ay2 |
| b | voiced-stop | labial | max | consonant |
| ch | stop | front | max | consonant |
| d | voiced-stop | alveolar | max | consonant |
| dh | voiced-fricative | dental | max | consonant |
| eh | vowel | mid-front | mid | eh |
| el | approximant | lateral | very-high | consonant |
| em | nasal | labial | max | consonant |
| en | nasal | alveolar | max | consonant |
| er | vowel | mid | mid | er |
| ey1 | vowel | front | mid-high | ey1 |
| ey2 | vowel | mid-front | high | ey2 |
| f | fricative | labial | max | consonant |
| g | voiced-stop | dorsal | max | consonant |
| hh | aspirated | unknown | max | consonant |
| ih | vowel | mid-front | high | ih |
| iy | vowel | front | very-high | iy |
| jh | voiced-stop | front | max | consonant |
| k | stop | dorsal | max | consonant |
| l | approximant | lateral | very-high | consonant |
| m | nasal | labial | max | consonant |
| n | nasal | alveolar | max | consonant |
| ng | nasal | dorsal | max | consonant |
| ow1 | vowel | back | mid | ow1 |
| ow2 | vowel | mid-back | high | ow2 |
| oy1 | vowel | back | mid-low | oy1 |
| oy2 | vowel | mid-front | high | oy2 |
| p | stop | labial | max | consonant |
| r | approximant | retroflex | mid-low | consonant |
| s | fricative | alveolar | max | consonant |
| sh | fricative | front | max | consonant |
| t | stop | alveolar | max | consonant |
| th | fricative | dental | max | consonant |
| uh | vowel | mid-back | high | uh |
| uw | vowel | back | very-high | uw |
| v | voiced-fricative | labial | max | consonant |
| w | approximant | back | very-high | consonant |
| y | approximant | front | very-high | consonant |
| z | voiced-fricative | alveolar | max | consonant |
| zh | voiced-fricative | front | max | consonat |

# Bibliography

Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2-3):109–118.

Alku, P. (2011). Glottal inverse filtering analysis of human voice production—a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36(5):623–650.

Alzheimer's Association (2017). 2017 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 13(4):325–373.

An, G., Brizan, D. G., Ma, M., Morales, M., Syed, A. R., and Rosenberg, A. (2015). Automatic recognition of unified Parkinson's disease rating from speech with acoustic, i-vector and phonotactic features. In *Proceedings of Interspeech*.

Aradilla, G., Vepa, J., and Bourlard, H. (2007). An acoustic model based on kullback-leibler divergence for posterior features. In *Proceedings of ICASSP*, volume 4, pages IV–657.

Arias-Vergara, T. et al. (2017). Parkinson's disease and aging: analysis of their effect in phonation and articulation of speech. *Cognitive Computation*, 9(6):731–748.

Arjmandi, M. K. and Pooyan, M. (2012). An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. *Biomedical signal processing and control*, 7(1):3–19.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).

Berus, L., Klancnik, S., Brezocnik, M., and Ficko, M. (2018). Classifying parkinson's disease based on acoustic measures using artificial neural networks. *Sensors*, 19(1):16.

Bhati, S., Velazquez, L. M., Villalba, J., and Dehak, N. (2019). Lstm siamese network for parkinson's disease detection from speech. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE.

Blahut, R. E. (1974). Hypothesis testing and information theory. *IEEE Transactions on Information Theory*, IT-20(4):405–417.

Bloem, B. R., Okun, M. S., and Klein, C. (2021). Parkinson's disease. *The Lancet*, 397(10291):2284–2303.

Bocklet, T., Steidl, S., Nöth, E., and Skodda, S. (2013). Automatic evaluation of Parkinson's speech-acoustic, prosodic and voice related cues. In *Proceedings of Interspeech*, pages 1149–1153.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345.

Bratzke, D., Rolke, B., Ulrich, R., and Peters, M. (2007). Central slowing during the night. *Psychological Science*, 18(5):456–461.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a "siamese" time delay neural network. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems 6*, pages 737–744. Morgan-Kaufmann.

Carletta, J. et al. (2005). The AMI meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 28–39. Springer.

Cernak, M. et al. (2017). Characterisation of voice quality of Parkinson's disease using differential phonological posterior features. *Computer Speech & Language*, 46:196–208.

Chen, L.-W. and Rudnicky, A. (2021). Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition. *arXiv preprint arXiv:2110.06309*.

Chomsky, N. and Halle, M. (1968). *The Sound Patterns in English*. MIT Press.

Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V. S., Christensen, H., Blackburn, D., Schuller, B. W., Magimai-Doss, M., Strik, H., et al. (2020). A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition. In *Proceedings of Interspeech*, pages 2182–2186.

De Bodt, M. S., Hernández-Diaz Huici, M. E., and Van De Heyning, P. H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders*, 35(3):283–292.

De Bodt, M. S., Huici, M. E. H.-D., and Van De Heyning, P. H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of communication disorders*, 35(3):283–292.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Dibazar, A. A., Narayanan, S., and Berger, T. W. (2002). Feature analysis for automatic detection of pathological speech. In *Proceedings of the second joint 24th annual conference and the annual fall meeting of the biomedical engineering societyi engineering in medicine and biology*, volume 1, pages 182–183. IEEE.

Dorsey, E. R. et al. (2018). Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 17(11):939–953.

Dubagunta, S. P. and Magimai-Doss, M. (2019). Using speech production knowledge for raw waveform modelling based Styrian dialect identification. In *Proceedings of Interspeech*, pages 2383–2387.

Dubagunta, S. P., Vlasenko, B., and Magimai.-Doss, M. (2019). Learning voice source related information for depression detection. In *Proceedings of ICASSP*, pages 6525–6529.

Dubagunta, S. P., Vlasenko, B., and Magimai.-Doss, M. (2019). Learning voice source related information for depression detection. In *Proceedings of ICASSP*, pages 6525–6529.

Dubagunta, S. P., Vlasenko, B., and Magimai.-Doss, M. (2019). Learning voice source related information for depression detection. In *Proceedings of ICASSP*.

Duffy, J. R. (2012). *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Elsevier Health Sciences.

Elsner, D., Langer, S., Ritz, F., Mueller, R., and Illium, S. (2019). Deep neural baselines for computational paralinguistics. *Proceedings of Interspeech*, pages 2388–2392.

Enderby, P. M. and Palmer, R. (2008). *FDA-2: Frenchay Dysarthria Assessment: Examiner's Manual*. Pro-ed.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Falk, T. H., Chan, W.-Y., and Shein, F. (2012). Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication*, 54(5):622–631.

Falk, T. H., Hummel, R., and Chan, W. (2011). Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility. In *Proceedings of ICASSP*, pages 4480–4483.

Ferrier, L., Shane, H., Ballard, H., Carpenter, T., and Benoit, A. (1995). Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication*, 11(3):165–175.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini mental state a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., and Schuller, B. (2017). audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research*, 18(1):6340–6344.

Fritsch, J., Dubagunta, S. P., and Doss, M. M. (2020). Estimating the degree of sleepiness by integrating articulatory feature knowledge in raw waveform based CNNs. In *Proceedings of ICASSP*, pages 6534–6538. IEEE.

Fritsch, J. and Magimai-Doss, M. (2021). Utterance verification-based dysarthric speech intelligibility assessment using phonetic posterior features. *Ieee signal processing letters*, 28:224–228.

Fritsch, J., Wankerl, S., and Nöth, E. (2019). Automatic diagnosis of Alzheimer's disease using neural network language models. In *Proceedings of ICASSP*, pages 5841–5845.

Godino-Llorente, J., Shattuck-Hufnagel, S., Choi, J., Moro-Velázquez, L., and Gómez-García, J. (2017). Towards the identification of idiopathic parkinson's disease from the speech. new articulatory kinetic biomarkers. *PloS one*, 12(12):e0189583.

Goetz, C. G. et al. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement disorders*, 23(15):2129–2170.

Gosztolya, G. (2019). Using Fisher vector and bag-of-audio-words representations to identify Styrian dialects, sleepiness, baby & orca sounds. *Proceedings of Interspeech*, pages 2413–2417.

Hafner, M., Stepanek, M., Taylor, J., Troxel, W. M., and Stolk, C. V. (2016). Why sleep matters — the economic costs of insufficient sleep: A cross-country comparative analysis. www.rand.org/pubs/research_reports/RR1791.html, Accessed: 2019-10-20.

Halpern, B. M., Fritsch, J., Hermann, E., van Son, R., Scharenborg, O., and Magimai-Doss, M. (2021). An objective evaluation framework for pathological speech synthesis. In *14th ITG Conference*, pages 1–5. VDE.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). CNN architectures for large-scale audio classification. In *Proceedings of ICASSP*, pages 131–135.

Ho, A. K. et al. (1998). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural neurology*, 11(3):131–137.

Hönig, F., Batliner, A., Nöth, E., Schnieder, S., and Krajewski, J. (2014). Acoustic-prosodic characteristics of sleepy speech–between performance and interpretation. In *Proceedings Speech Prosody*, pages 864–868.

Hönig, F., Stemmer, G., Hacker, C., and Brugnara, F. (2005). Revising perceptual linear prediction (PLP). In *Ninth European Conference on Speech Communication and Technology*.

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2019a). Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. In *Proceedings of ICASSP*, pages 6405–6409. IEEE.

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2019b). Spectral subspace analysis for automatic assessment of pathological speech intelligibility. In *Proceedings of Interspeech*.

Jensen, J. and Taal, C. H. (2016). An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022.

Johnson, D., Ellis, D., Oei, C., Wooters, C., and Faerber, P. (2004). Quicknet. *www1.icsi.berkeley.edu/Speech/qn.html*.

Kabil, S. H., Muckenhirn, H., and Magimai-Doss, M. (2018). On learning to identify genders from raw speech signal using CNNs. In *Proceedings of Interspeech*, pages 287–291.

Kadiri, S. R. and Alku, P. (2019). Analysis and detection of pathological voice using glottal source features. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):367–379.

Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication*, 15(1):52–60.

Kant, H. (2011). *Lebenslauf, zweiter Absatz: Erzählungen*. Aufbau Digital.

Kent, R., Weismer, G., Kent, J., and Rosenbek, J. (1989). Toward phonetic intelligibility testing in dysarthria. *The Journal of speech and hearing disorders*, 54:482–99.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., and Frame, S. (2008). Dysarthric speech database for universal access research. In *Ninth Annual Conference of the International Speech Communication Association*.

Kim, M. J., Kim, Y., and Kim, H. (2015). Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):694–704.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klautau, A. (2001). ARPABET and the TIMIT alphabet. [Online; accessed 15th April 2023].

Knuijt, S., Kalf, J. G., van Engelen, B. G., de Swart, B. J., and Geurts, A. C. (2017). The Radboud dysarthria assessment: development and clinimetric evaluation. *Folia Phoniatrica et Logopaedica*, 69(4):143–153.

Koo, J., Lee, J. H., Pyo, J., Jo, Y., and Lee, K. (2020). Exploiting multi-modal features from pretrained networks for Alzheimer's dementia recognition. *arXiv preprint arXiv:2009.04070*.

Korzekwa, D., Lorenzo-Trueba, J., Drugman, T., Calamaro, S., and Kostek, B. (2021). Weakly-supervised word-level pronunciation error detection in non-native english speech. *arXiv preprint arXiv:2106.03494*.

Kostyk, B. E. and Rochet, A. P. (1998). Laryngeal airway resistance in teachers with vocal fatigue: A preliminary study. *Journal of Voice*, 12(3):287–299.

Koutrouvelis, A. I., Kafentzis, G. P., Gaubitch, N. D., and Heusdens, R. (2015). A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):316–328.

Krajewski, J. et al. (2010). Detecting fatigue from steering behaviour applying continuous wavelet transform. In *Proceedings of Measuring Behaviour*, pages 326–329.

Ladefoged, P. (1993). *A Course in Phonetics*. Harcourt Brace College Publishers.

Ladefoged, P. and Johnson, K. (2014). *A course in phonetics*. Cengage learning.

Legendre, S., Liss, J., and Lotto, A. (2009). Discriminating dysarthria type and predicting intelligibility from amplitude modulation spectra. *The Journal of the Acoustical Society of America*, 125:2530.

Li, R., Wu, Z., Jia, J., Zhao, S., and Meng, H. (2019). Dilated residual network with multi-head self-attention for speech emotion recognition. In *Proceedings of ICASSP*, pages 6675–6679. IEEE.

Liu, Y., Zhao, W.-L., Ngo, C.-W., Xu, C.-S., and Lu, H.-Q. (2010). Coherent bag-of audio words model for efficient large-scale video copy detection. In *Proceedings of the ACM international conference on image and video retrieval*, pages 89–96.

Logemann, J. A. et al. (1978). Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *Journal of Speech and hearing Disorders*, 43(1):47–57.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. In *Proceedings Interspeech*, Shanghai, China.

Martínez, D., Lleida, E., Green, P., Christensen, H., Ortega, A., and Miguel, A. (2015). Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Transactions on Accessible Computing*, 6(3):1–21.

Middag, C. (2012). *Automatic analysis of pathological speech*. PhD thesis, Ghent University.

Middag, C., Martens, J.-P., Van Nuffelen, G., and De Bodt, M. (2009). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing*, 2009(1):1–9.

Montaña, D., Campos-Roca, Y., and Pérez, C. J. (2018). A Diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of Parkinson's disease. *Computer methods and programs in biomedicine*, 154:89–97.

Moro-Velazquez, L. et al. (2019). Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's Disease. *Scientific reports*, 9(1):1–16.

Muckenhirn, H., Magimai.-Doss, M., and Marcel, S. (2017). End-to-end convolutional neural network-based voice presentation attack detection. In *International Joint Conference on Biometrics*.

Muckenhirn, H., Magimai-Doss, M., and Marcell, S. (2018). Towards directly modeling raw speech signal for speaker verification using cnns. In *Proceedings of ICASSP*, pages 4884–4888. IEEE.

Murty, K. S. R. and Yegnanarayana, B. (2008). Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1602–1613.

Nallanthighal, V. S., Mostaani, Z., Härmä, A., Strik, H., and Magimai-Doss, M. (2021). Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings. *Neural Networks*, 141:211–224.

Narendra, N. P. and Alku, P. (2020). Glottal source information for pathological voice detection. *IEEE Access*, 8:67745–67755.

Norel, R. et al. (2020). Speech-based characterization of dopamine replacement therapy in people with Parkinson's disease. *NPJ Parkinson's disease*, 6(1):1–8.

Noroozi, F., Sapiński, T., Kamińska, D., and Anbarjafari, G. (2017). Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*, 20(2):239–246.

Novotnỳ, M. et al. (2020). Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with Parkinson's disease: Correlation between acoustic speech characteristics and non-speech motor performance. *Biomedical Signal Processing and Control*, 57:101818.

Orozco-Arroyave, J. R. (2016). *Analysis of speech of people with Parkinson's disease*, volume 41. Logos Verlag Berlin GmbH.

Orozco-Arroyave, J. R. et al. (2014a). New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 342–347.

Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Skodda, S., Rusz, J., and Nöth, E. (2014b). Automatic detection of parkinson's disease from words uttered in three different languages. In *Proceedings of Interspeech*.

Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Skodda, S., Rusz, J., and Nöth, E. (2015). Voiced/unvoiced transitions in speech as a potential bio-marker to detect parkinson's disease. In *Proceedings of Interspeech*.

Paja, M. S. and Falk, T. H. (2012). Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech. In *Proceedings of Interspeech*.

Palaz, D., Collobert, R., and Magimai-Doss, M. (2013). Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. *Proceedings of Interspeech*, pages 1766–1770.

Palaz, D., Magimai.-Doss, M., and Collobert, R. (2019). End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. *Speech Communication*, 108:15–32.

Pan, Y., Mirheidari, B., Reuber, M., Venneri, A., Blackburn, D., and Christensen, H. (2019). Automatic hierarchical attention neural network for detecting ad. In *Proceedings of Interspeech*, pages 4105–4109.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543, Doha, Qatar. ACL.

Pereira da Silva, A., Feliciano, T., Vaz Freitas, S., Esteves, S., and Almeida e Sousa, C. (2015). Quality of life in patients submitted to total laryngectomy. *Journal of Voice*, 29(3):382–388.

Quintas, S., Mauclair, J., Woisard, V., and Pinquier, J. (2022). Automatic assessment of speech intelligibility using consonant similarity for head and neck cancer. In *Interspeech 2022*.

Ramig, L. O. and Verdolini, K. (1998). Treatment efficacy: voice disorders. *Journal of Speech, Language, and Hearing Research*, 41(1):S101–S116.

Rasipuram, R., Cernak, M., Nanchen, A., and Magimai.-Doss, M. (2015). Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities. In *Proceedings of Interspeech*.

Rasipuram, R. and Magimai.-Doss, M. (2016). Articulatory feature based continuous speech recognition using probabilistic lexical modeling. *Computer Speech & Language*, 36:233–259.

Ravanelli, M. and Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028.

Ravi, V., Park, S. J., Afshan, A., and Alwan, A. (2019). Voice quality and between-frame entropy for sleepiness estimation. *Proceedings of Interspeech*, pages 2408–2412.

Rudzicz, F., Namasivayam, A. K., and Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46(4):523–541.

Rusz, J., Cmejla, R., Ruzickova, H., and Ruzicka, E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *The journal of the Acoustical Society of America*, 129(1):350–367.

Rusz, J. et al. (2013). Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task. *The Journal of the Acoustical Society of America*, 134(3):2171–2181.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-26(1):43–49.

Scheltens, P., Blennow, K., Breteler, M. M., De Strooper, B., Frisoni, G. B., Salloway, S., and Van der Flier, W. M. (2016). Alzheimer's disease. *The Lancet*, 388(10043):505–517.

Schmitt, M. and Schuller, B. (2017). openXBOW–introducing the Passau open-source cross-modal bag-of-words toolkit. *Journal of Machine Learning Research*, 18(96):1–5.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. In *Proceedings of Interspeech*.

Schuller, B., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., Vollmann, R., Roelen, S.-D., Schnieder, S., Bergelson, E., et al. (2019). The Interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity. In *Proceedings of Interspeech*.

Schuller, B. et al. (2014). Medium-term speaker states—a review on intoxication, sleepiness and the first challenge. *Computer Speech & Language*, 28(2):346–374.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al. (2013). The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings of Interspeech*.

Schuller, B. W., Batliner, A., Bergler, C., Mascolo, C., Han, J., Lefter, I., Kaya, H., Amiriparian, S., Baird, A., Stappen, L., et al. (2021). The Interspeech 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates. In *Proceedings of Interspeech*, pages 431–435.

Schuller, B. W., Batliner, A., Bergler, C., Messner, E.-M., Hamilton, A., Amiriparian, S., Baird, A., Rizos, G., Schmitt, M., Stappen, L., et al. (2020). The Interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks. In *Proceedings of Interspeech*.

Schuster, M., Maier, A., Haderlein, T., Nkenke, E., Wohlleben, U., Rosanowski, F., Eysholdt, U., and Nöth, E. (2006). Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition. *International Journal of Pediatric Otorhinolaryngology*, 70(10):1741–1747.

Shahid, A., Wilkinson, K., Marcu, S., and Shapiro, C. M. (2011). Karolinska sleepiness scale (KSS). In *STOP, THAT and One Hundred Other Sleep Scales*, pages 209–210. Springer.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *Proceedings of ICASSP*, pages 4779–4783.

Skodda, S., Visser, W., and Schlegel, U. (2011). Vowel articulation in Parkinson's disease. *Journal of voice*, 25(4):467–472.

Smith, L. K. and Goberman, A. M. (2014). Long-time average spectrum in individuals with Parkinson disease. *NeuroRehabilitation*, 35(1):77–88.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: robust dnn embeddings for speaker recognition. In *Proceedings of ICASSP*, pages 5329–5333.

Stemmer, G., Nöth, E., and Parsa, V. (2010). Atypical speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010:1–2.

Story, B. H. (2002). An overview of the physiology, physics and modeling of the sound source for vowels. *Acoustical Science and Technology*, 23(4):195–206.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136.

Tanaka, Y., Nishio, M., and Niimi, S. (2011). Vocal acoustic characteristics of patients with Parkinson's disease. *Folia Phoniatrica et logopaedica*, 63(5):223–230.

Tanner, K., Roy, N., Ash, A., and Buder, E. H. (2005). Spectral moments of the long-term average spectrum: sensitive indices of voice change after therapy? *Journal of Voice*, 19(2):211–222.

Titze, I. R. and Martin, D. W. (1998). Principles of voice production.

Travieso, C. M. et al. (2017). Detection of different voice diseases based on the nonlinear characterization of speech signals. *Expert Systems with Applications*, 82:184–195.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of ICASSP*, pages 5200–5204.

Ullmann, R., Magimai.-Doss, M., and Bourlard, H. (2015a). Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences. In *Proceedings of ICASSP*.

Ullmann, R., Rasipuram, R., Magimai.-Doss, M., and Bourlard, H. (2015b). Objective intelligibility assessment of text-to-speech systems through utterance verification. In *Proceedings of Interspeech*.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. In *9th ISCA Speech Synthesis Workshop*, pages 125–125.

Van Nuffelen, G., De Bodt, M., Middag, C., and Martens, J.-P. (2009). Dutch corpus of pathological and normal speech (COPAS). *Antwerp University Hospital and Ghent University, Tech. Rep*.

van Son, R. J., Middag, C., and Demuynck, K. (2018). Vowel space as a tool to evaluate articulation problems. In *Proceedings of Interspeech*, volume 2018, pages 357–361.

Vásquez-Correa, J., Klumpp, P., Orozco-Arroyave, J. R., and Nöth, E. (2019). Phonet: a tool based on gated recurrent neural networks to extract phonological posteriors from speech. *Proceedings of Interspeech*, pages 549–553.

Vasquez-Correa, J. C., Arias-Vergara, T., Schuster, M., Orozco-Arroyave, J. R., and Nöth, E. (2020). Parallel representation learning for the classification of pathological speech: studies on parkinson's disease and cleft lip and palate. *Speech Communication*, 122:56–67.

Vásquez-Correa, J. C., Fritsch, J., Orozco-Arroyave, J. R., Nöth, E., and Magimai-Doss, M. (2021). On modeling glottal source information for phonation assessment in Parkinson's disease. In *Proceedings of Interspeech*, pages 26–30.

Vásquez-Correa, J. C., Orozco-Arroyave, J. R., Bocklet, T., and Nöth, E. (2018). Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. *Journal of communication disorders*, 76:21–36.

Vásquez-Correa, J. C., Orozco-Arroyave, J. R., and Nöth, E. (2017a). Convolutional neural network to model articulation impairments in patients with Parkinson's disease. In *Proceedings of Interspeech*, pages 314–318.

Vásquez-Correa, J. C., Serra, J., Orozco-Arroyave, J. R., Vargas-Bonilla, J. F., and Nöth, E. (2017b). Effect of acoustic conditions on algorithms to detect parkinson's disease from speech. In *Proceedings of ICASSP*, pages 5065–5069.

Villatoro-Tello, E., Dubagunta, S. P., Fritsch, J., Motlicek, P., and Magimai-Doss, M. (2021). Late fusion of the available lexicon and raw waveform-based acoustic modeling for depression and dementia recognition. In *Proceedings of Interspeech*, pages 1927–1931.

Vogel, A. P., Fletcher, J., and Maruff, P. (2010). Acoustic analysis of the effects of sustained wakefulness on speech. *The Journal of the Acoustical Society of America*, 128(6):3747–3756.

Wei, X., Cucchiarini, C., van Hout, R., and Strik, H. (2022). Automatic speech recognition and pronunciation error detection of dutch non-native speech: cumulating speech resources in a pluricentric language. *Speech Communication*, 144:1–9.

Wikimedia (2023). Speech production: Human vocal apparatus used to produce speech.

Witt, S. M. and Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3):95–108.

World Health Organization (2018). Towards a dementia plan: A WHO guide. WHO. 82 pages.

Wu, H., Wang, W., and Li, M. (2019a). The DKU-Lenovo systems for the Interspeech 2019 computational paralinguistic challenge. *Proceedings of Interspeech*, pages 2433–2437.

Wu, P., Rallabandi, S., Black, A. W., and Nyberg, E. (2019b). Ordinal triplet loss: Investigating sleepiness detection from speech. *Proceedings of Interspeech*, pages 2403–2407.

Yang, S., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., and Guan-Ting, e. a. (2021). SUPERB: Speech processing universal performance benchmark. In *Proceedings of Interspeech*, pages 1194–1198.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT 2016*, pages 1480–1489, San Diego, California, USA.

Yeh, S.-L. et al. (2019). Using attention networks and adversarial augmentation for Styrian dialect continuous sleepiness and baby sound recognition. *Proceedings of Interspeech*, pages 2398–2402.

Yorkston, K. M., Strand, E. A., and Kennedy, M. R. (1996). Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology (ASHA)*, 5(1):55–66.

Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease. In *Proceedings of Interspeech*, volume 2020, pages 2162–6.

Zhang, L., Zhao, Z., Ma, C., Shan, L., Sun, H., Jiang, L., Deng, S., and Gao, C. (2020). End-to-end automatic pronunciation error detection based on improved hybrid ctc/attention architecture. *Sensors*, 20(7):1809.

# Julian Fritsch

julian.fritsch@epfl.ch | idiap.ch/~jfritsch | +41767227035 | Nationality: German

## Areas of Interest and Skills

Speech processing, machine learning, deep learning.

## Education

**Idiap Research Institute, Switzerland**
*Research Assistant: Speech processing, deep learning* *2018 – 2022*

**École polytechnique fédérale de Lausanne (EPFL), Switzerland**
*PhD Student. Thesis: Interpretable speech pathology detection* *2018 – 2022*
*Detection and analysis of phonetic differences in atypical speech for early detection or therapy*

**FAU Erlangen-Nuremberg, Germany**
*Master of Science in Medical Engineering: Medical image and data processing* *2015 – 2018*
*Thesis: Language models for dementia detection*

**FAU Erlangen-Nuremberg, Germany**
*Bachelor of Science in Medical Engineering: Biomedical signal analysis, pattern recognition* *2010 – 2015*
*Thesis: Automatic speech recognition with the Kaldi toolkit*

## Work Experience

**Working Student** May–Aug 2017
*E&L medical systems GmbH, Germany*

- Prototyping a German online ASR system with Kaldi to navigate in a medical documentation software

**Research Intern** Nov 2015 – Mar 2016
*Eriksholm Research Center, Denmark*

- Implementing a hearing aid acoustic feedback elimination algorithm in Assembler

## Research Experience

**Research Intern** Nov 2016 – Mar 2017
*International Audio Laboratories, Germany*

- Applying neural networks to bandwidth extension

**Student Research Assistant** Jan–Sep 2015
*Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany*

- Recognition experiments on audio data from dementia patients
- Transcription of an algorithm for pitch tracking in speech from Python to JavaScript

**Student Research Assistant** 2013 – 2015
*Chair of Information Transmission, FAU Erlangen-Nuremberg, Germany*

- Matlab tutoring

## Technical Skills

**Languages**: Python, Bash, Matlab
**Developer Tools**: Git, Vim, VS Code
**Libraries**: NumPy, Scikit-learn, Matplotlib
**Deep Learning Libraries**: Keras, Tensorflow

## Languages

**German**: Native profiency
**English**: Full professional proficiency
**French**: Full professional proficiency