EPFL

Thèse n°9483

Automatic pathological speech assessment

Présentée le 13 juin 2022

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

Parvaneh JANBAKHSHI

Acceptée sur proposition du jury

Dr J.-M. Vesin, président du jury
Prof. H. Bourlard, Dr I. Kodrasi, directeurs de thèse
Prof. Ph. Green, rapporteur
Prof. M. G. Christensen, rapporteur
Prof. J.-Ph. Thiran, rapporteur

École
polytechnique
fédérale
de Lausanne

2022

Abstract

Many pathologies cause impairments in the speech production mechanism resulting in reduced speech intelligibility and communicative ability. To assist the clinical diagnosis, treatment and management of speech disorders, automatic pathological speech assessments are indispensable. Such automatic assessments provide reliable, objective, and cost-effective assessment in contrast to subjective and time-consuming auditory-perceptual analyses performed by clinicians. Among crucial automatic analyses for developing potential computer-aided tools are speech pathology detection, i.e., discriminating between normal and pathological speech, and speech intelligibility assessment, i.e., predicting an intelligibility index correlated with the percentage of words correctly understood by human listeners. The goal of this thesis is to propose novel data-driven approaches to aid the development of a clinical assistive tool for automatic pathological speech assessment with two purposes, i.e., pathological speech detection and intelligibility assessment.

First, we focus on the development of novel machine learning approaches to address the pathological speech detection task. Motivated by the clinical evidence on spectro-temporal distortions associated with pathological speech, we propose a subspace-based speech pathology detection approach that relies on analyzing subspaces spanned by the dominant spectral or temporal patterns of speech. Although the temporal subspace-based approach yields a high performance, it requires time-alignment and having access to phonetically-balanced utterances from all speakers. To avoid the time-alignment and also to assess the efficacy of deep learning approaches for such a task, we propose analyzing pairwise distance matrices computed from speech representations using convolutional neural networks. Furthermore, to be able to achieve pathological speech detection without requiring constraints on the phonetic content, we propose different supervised representation learning approaches using convolutional neural networks to learn robust and relevant feature representations. We demonstrate the effectiveness of the proposed approaches through different experiments across different databases.

Second, we focus on developing reliable automatic pathological speech intelligibility measures overcoming several drawbacks of the state-of-the-art measures while outperforming them. We first propose a measure based on short-time objective intelligibility assessment. Further, we provide a solution to ensure its applicability across scenarios with different phonetic content across speakers. We also propose intelligibility measures based on analyzing speech subspaces. The subspace-based intelligibility measures are applicable to different scenarios while overcoming the drawbacks of the previously described measure. We validate the performance

of the proposed measures across languages and diseases.
Finally, insights are provided on a potential clinical assistive tool for pathological speech detection and intelligibility assessment. To this end, we jointly validate the applicability of two of the previously described approaches, i.e., temporal subspace-based speech pathology detection and short-time objective intelligibility assessment. As our approaches for both tasks achieve a high performance independently of the language and disease, we confirm the possibility of developing such a multi-purpose clinical assistive tool.

Keywords: pathological speech intelligibility, pathological speech detection, ESTOI, convolutional neural network, subspace-based learning, supervised speech representation learning, feature separation, dysarthria, Parkinson's disease, Cerebral Palsy, Amyotrophic Lateral Sclerosis, hearing impairment

Résumé

De nombreuses pathologies causent des troubles dans le mécanisme de production de la parole qui résultent en une réduction de l'intelligibilité vocale et de la capacité à communiquer. Afin d'aider au diagnostic clinique, au traitement et à la prise en charge des troubles du langage, les évaluations automatiques de la parole pathologique sont indispensables. De telles analyses automatiques offrent une évaluation fiable, objective et rentable, en contraste avec des analyses audio-perceptives subjectives et longues réalisées par des médecins. Parmi les évaluations automatiques principales pour le développement de potentiels outils aidés par ordinateur se trouvent les approches de détection de pathologies de la parole, i.e. distinguer la parole normale de la parole pathologique, et les approches d'évaluation de l'intelligibilité de la parole, i.e. prédire un indice d'intelligibilité corrélé avec le pourcentage de mots correctement compris par des auditeurs humains. L'objectif de cette thèse est de proposer des approches novatrices basées sur les données pour aider au développement d'un outil clinique d'assistance à l'évaluation automatique de la parole pathologique avec deux buts : la détection de la parole pathologique, et l'évaluation de l'intelligibilité de la parole pathologique.

En premier lieu, nous nous concentrons sur le développement de nouvelles approches d'apprentissage machine pour résoudre la tâche de détection de la parole pathologique. Motivés par des preuves cliniques de distorsions spectro-temporelles associées à la parole pathologique, nous proposons une approche subspatiale de détection de la parole pathologique qui repose sur l'analyse de sous-espaces couverts par les motifs spectraux ou temporels dominants de la parole. Malgré que l'approche basée sur les sous-espaces temporels obtient une haute performance, elle nécessite un alignement temporel ainsi que l'accès à des échantillons phonétiquement équilibrés de tous les sujets d'étude. Pour éviter l'alignement temporel et aussi évaluer l'efficacité de l'apprentissage profond pour une telle tâche, nous proposons une approche basée sur l'analyse de matrices de distances paire-à-paire calculées à partir de représentations de la parole utilisant des réseaux de neurones convolutionnels. De plus, afin de pouvoir réaliser la détection de la parole pathologique sans nécessiter de contraintes sur le contenu phonétique, nous proposons différentes approches d'apprentissage supervisé utilisant des réseaux de neurones convolutionnels pour apprendre des représentations des données robustes et pertinentes pour la détection de la parole pathologique. Nous démontrons l'efficacité des approches proposées par différentes expériences sur différentes bases de données.

En second lieu, nous nous concentrons sur le développement de mesures automatiques fiables de l'intelligibilité de la parole pathologique surmontant plusieurs inconvénients des méthodes

de pointe tout en surpassant leurs performances. Nous proposons d'abord une mesure basée sur une évaluation objective à court terme de l'intelligibilité. Ensuite, nous fournissons une méthode pour assurer son applicabilité à de multiples scénarios avec des contenus phonétiques différents parmi les sujets. Nous proposons aussi des mesures d'intelligibilité basées sur l'analyse de sous-espaces de la parole. Les mesures d'intelligibilité basées sur les sous-espaces sont applicables dans différents scénarios tout en résolvant les inconvénients de la méthode décrite précédemment. Nous validons la performance et la capacité de généralisation des mesures proposées sur diverses langues et pathologies.
Finalement, nous fournissons des perspectives pour un potentiel outil clinique d'assistance à la détection de la parole pathologique et à l'évaluation de l'intelligibilité. À cette fin, nous validons conjointement l'applicabilité de deux des approches décrites précédemment, i.e. la détection de pathologies de la parole basée sur les sous-espaces temporels et l'évaluation objective à court terme de l'intelligibilité. Au vu de la performance élevée obtenue par nos méthodes sur ces deux tâches indépendamment de la langue et de la pathologie, nous confirmons la possibilité de développer un tel outil d'assistance clinique polyvalent.

Mots-clés : intelligibilité de la parole pathologique, détection de la parole pathologique, intelligibilité objective à court terme, réseaux de neurones convolutionnels, apprentissage sur sous-espaces, représentation supervisée de la parole, séparation des caractéristiques, dysarthrie, maladie de Parkinson, infirmité motrice cérébrale, sclérose latérale amyotrophique, déficience auditive

Acknowledgements

I would like to express my thanks and gratitude to my thesis supervisors. I first thank Prof. Hervé Bourlard for giving me the opportunity to work with him and for his guidance, insights, and motivational comments throughout my PhD. I also thank Dr. Ina Kodrasi, who constantly mentored me throughout my PhD and gave me a great deal of support and assistance, without whom this thesis would not have been possible. I know her as an exceptionally intelligent researcher and also a patient mentor from whom I could not learn enough academic and personal skills. I find her intelligence, independence, and perseverance inspiring for women in science such as myself.

I would like to thank the jury members of my thesis, i.e., Dr. Jean-Marc Vesin, Prof. Jean-Philippe Thiran, Prof. Philip Green, and Prof. Mads Græsbøll Christensen for their insightful questions and comments. I would also like to thank the Swiss National Science Foundation for funding my PhD research through the MoSpeeDi project.

I am also deeply grateful to the administrative team and help-desk at Idiap Research Institute for their invaluable support with technical and non-technical resources.

There is a saying "a journey is best measured in friends, rather than miles" and I think such a saying is also very applicable to all of us who started our PhD journeys far from our homelands and our families. I never imagined that my PhD journey could allow me to meet the kindest people and to shape some of the strongest friendships in my life. I am very lucky and grateful to have my friends at Idiap who have been there for me through good and bad times. I especially thank my genuine squad, i.e., Suraj Srinivas, Sargam Vyas, and Suhan Shetty for being my family away from my family all these years. I also thank many other friends at Idiap whom I did not hesitate to seek advice and help from and whom I shared countless amazing discussions, tea, dinner, ski trips, and hikes.

Finally, last but by no means least, I am most grateful to my whole family for their support, their faith and trust in me, and the freedom they provided me to pursue my interests.

Martigny, April 21, 2022 Parvaneh Janbakhshi

Contents

List of Figures

List of Tables

1 Introduction

Speech is a very complex activity which requires the synchronous contraction of many muscle groups associated with respiration, laryngeal function, airflow direction, and articulation. As a result, speech can be impaired in different ways because of different pathologies caused by genetic influences, physical deformities, hearing loss, or neurological malfunctions. For example, dysarthria of speech is a common neurological speech impairment known as motor speech disorder which results from disturbances of the muscular control on the movement mechanism necessary for the execution of speech (Duffy, 2000). Dysarthria arises in several neurological etiologies, e.g., stroke, Cerebral Palsy (CP), Amyotrophic Lateral Sclerosis (ALS), and Parkinson's Disease (PD) (Darley et al., 1969; Flipsen and Parker, 2008). Depending on the origin and the severity of the speech impairment, several components of the speech production mechanism can be affected such as respiration, phonation, resonance, and articulation, yielding an abnormal quality of speech as well as reduced intelligibility and communicative ability (Enderby, 2013).

Monitoring changes in speech is crucial for providing an accuracte clinical diagnosis of speech pathologies and therapeutic feedback, since speech changes reveal important information about the pathology and its severity. In addition, in case of progressive neurologic conditions, speech analysis can provide an early evidence of the neurological disease evolution (Duffy, 2000). To monitor speech changes, speech pathologists use auditory-perceptual evaluation in a range of speech tasks and listening paradigms. One component of the auditory-perceptual assessment is the evaluation of specific perceptual cues associated with speech traits distorted by pathologies, such as roughness, breathiness, or nasality (Sussman and Tjaden, 2012). These evaluations are then used for pathology detection, i.e., discrimination between normal and pathological speech. Such clinical assessments for pathological speech diagnosis are time-consuming and expensive for screening a large number of subjects, and the results of these analyses may be inconsistent due to many factors such as different subjective opinions between clinicians depending on the clinicians' experience, different types of rating scales, and different speech tasks under study (Oates, 2009; Baghai-Ravary and Beet, 2012). Furthermore, such clinical perceptual assessments might not be able to detect a pathological condition at an

early stage (Gavidia-Ceballos and Hansen, 1996). After establishing the acoustic pathological characteristics elicited from the speech of the patients, another component of the clinical auditory-perceptual evaluation is required to determine the overall functional oral communicative performance, i.e., speech intelligibility assessment, which helps to characterize the severity of the speech pathology (Sussman and Tjaden, 2012). Impaired speech intelligibility can be a barrier to social engagement and education, which can affect the self-esteem and the quality of life for a patient. Therefore, intelligibility assessment guides speech therapy interventions aiming at improving speech intelligibility. The gold standard pathological speech intelligibility measure is based on subjective listening tests evaluating the percentage of words correctly understood by human listeners (Sussman and Tjaden, 2012; Landa et al., 2014). Such subjective assessments of intelligibility are labor-intensive, costly, and are also affected by the listener's familiarity with the patient's speech pathology and by the contextual/linguistic cues available in connected speech (Landa et al., 2014).

To assist the clinical diagnosis of speech pathologies and to avoid the drawbacks associated with clinical assessments, automatic pathological speech detection methods based on machine learning and signal processing can be used (Baghai-Ravary and Beet, 2012). Furthermore, as an efficient and economical substitute to subjective intelligibility assessment, automatic pathological speech intelligibility measures have been proposed (Maier et al., 2009; Middag et al., 2010; Bocklet et al., 2012; Martínez et al., 2015; Imed et al., 2017; Kalita et al., 2018). In contrast to the auditory-perceptual evaluation performed by clinicians, such automatic pathological speech analyses aiming at pathological speech detection and intelligibility assessment offer frequent, efficient, economical, and objective assessment tools. These techniques not only pave the way for more reliable and repeatable pathological speech analysis to be used for early diagnosis and disease management, but can also be used in speech therapy with the capability of being performed remotely (Wallen and Hansen, 1996; Baghai-Ravary and Beet, 2012).

Machine learning and signal processing techniques as a cross-disciplinary approach to assess pathological speech face several challenges. The performance of these techniques is affected by factors such as a) the broad range of pathologies causing speech impairments, b) the broad spectrum of impairments within a single pathological condition based on the disease severity, and c) several sources of noise which can be categorized into 3 subgroups i.e., c-1) inter-speaker variability referring to variabilities such as speaker-specific traits (e.g., speaker identity), language, or accent, c-2) intra-speaker variability such as fatigue, and c-3) variable recording conditions. Hence, developing a system which accurately models each paralinguistic aspect of speech is difficult and, due to the scarcity of sufficiently comprehensive databases, remains one of the most challenging tasks to date (Baghai-Ravary and Beet, 2012; Gupta et al., 2016).

This thesis presents approaches toward a clinical automated tool for pathological speech, aiming at simultaneous pathological speech detection and pathological speech intelligibility assessment. Such a clinical tool should ideally have a high agreement with human decisions

regarding diagnosis and intelligibility assessment, while offering objective and cost-effective acoustic analysis to further assist the clinical monitoring and management of speech disorders. Without limiting the conducted speech assessment in this thesis to a specific disease, we aim to develop approaches applicable to different atypicalities in speech. The pathological conditions we focus on are ALS, PD, CP, and speech disorders caused by hearing impairment (HI). At first, we focus on each component of the clinical tool separately and consider pathological speech detection and intelligibility assessment as two different tasks. Automatic intelligibility assessment mainly deals with quantifying acoustic characteristics associated with speech perception while pathology detection can not only rely on impaired perceptible dimensions of speech, especially for mild or early-stage pathological conditions. Considering the first component of the clinical tool, we propose novel approaches for automatic pathological speech detection which overcome many drawbacks of state-of-the-art methods and are generalisable across languages and diseases. Considering the second component of the clinical tool, we also propose reliable objective pathological speech intelligibility measures which are applicable to different scenarios and outperform current state-of-the-art intelligibility measures. Finally, by selecting advantageous methods among our proposed approaches for each task, we investigate the possibility of developing an automated tool for jointly evaluating both pathological speech assessment aspects to assist clinical speech screenings.

1.1 Motivations and contributions

The high-level objective of this thesis is to devise novel approaches to be used in an automatic pathological speech assessment tool for speech pathology detection and speech intelligibility assessment. More specifically, the main motivations and contributions of this thesis can be summarized as follows:

1. Current approaches aimed at automatic speech pathology detection are largely dominated by assessing the voice quality dimension of speech which is prominent in disorders associated with abnormal vocal fold function. Therefore, such approaches rely on the availability of controlled and somewhat less natural sustained phonation data. However, impairments such as dysarthria significantly affect other dimensions of speech as well, e.g., articulation dynamics which can be elicited from connected speech representative of daily voice. Automatic approaches exploiting connected speech analysis rely on classical machine learning approaches requiring handcrafting large-scale bruteforced acoustic feature sets, voice/unvoiced speech segmentation that is not robust when analyzing pathological speech, and acoustic modeling derived from other speech applications without being fully optimized for the speech pathology detection task. In addition, recently there has been a growing interest in the research community to leverage pure data-driven deep learning approaches. However, such approaches face challenges due to typically limited pathological training data. Therefore, the performance of many state-of-the-art approaches and also their generalizability across

languages and pathologies can be limited. In this thesis, we propose novel machine learning and deep learning approaches to tackle such drawbacks in analyzing connected speech using minimal prior knowledge, and we demonstrate their effectiveness across languages and diseases. First, motivated by the clinical evidence on spectro-temporal distortions associated with pathological speech, we propose a novel approach based on analyzing the dominant spectro-temporal patterns of healthy and pathological speech using a subspace-based learning technique. Then, by focusing on under-explored deep learning frameworks in this field, we propose a pairwise distance-based convolutional neural network which is motivated by advantages of pairwise training when limited training data is available. Further, motivated by the fact that the presence of non-relevant speaker variabilities in feature representations, e.g., speaker identity cues, can degrade the performance of pathological speech detection systems while learning optimal abstract features specific to the task can improve the performance, we propose methods to supervise convolutional neural networks training to learn more robust and relevant abstract acoustic cues for this task. The superiority of the proposed frameworks is then demonstrated by outperforming their counterpart baseline systems.

2. Among many approaches that have been proposed for automatic pathological speech intelligibility assessment, approaches exploiting healthy (i.e., perfectly intelligible) speech signals as references have shown superior performance. However, they are usually complex and require a large number of healthy speech recordings for training which limits their application for low-resource languages. In this thesis, we propose a reliable, robust, and simple objective intelligibility measure outperforming state-of-the-art measures. Developing intelligibility measures applicable to scenarios with fewer or no constraints on the phonetic content of speech used for training or evaluation is an under-explored topic in the field. Hence, aiming at extending the applicability of the intelligibility measure for different scenarios with variable constraints on the phonetic content of speech, we further propose more flexible intelligibility measures and demonstrate their generalisability across languages and diseases. All our measures are based on developing a single feature correlated with subjective intelligibility ratings, hence it is advantageous over methods that require many acoustic features (since training, and hence, overfitting is avoided).

3. By selecting the best performing and robust techniques among our proposed approaches for speech pathology detection and intelligibility assessment, we jointly evaluate the two tasks as a step toward a multi-purpose clinical tool for automatic pathological speech assessment.

1.2 Thesis outline

A schematic overview of the thesis is presented in Fig 1.1. The thesis is organized as follows.

Chapter 2 provides a general overview of the state-of-the-art automatic pathological speech

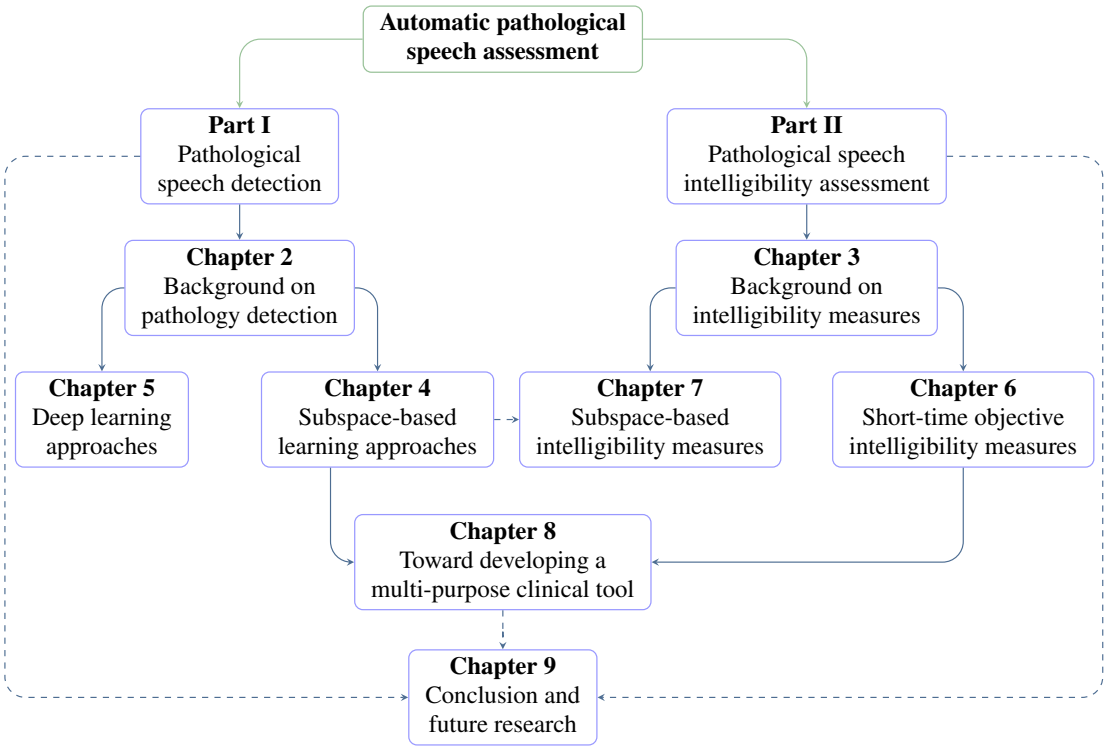


Figure 1.1 – Schematic overview of the thesis.

detection approaches. In addition, an overview of the data set information and protocols used in our thesis for pathological speech detection task is provided.

Chapter 3 presents a general overview of the state-of-the-art automatic pathological speech intelligibility measures. In addition, an overview of the data set information and protocols used in our thesis for intelligibility assessment is provided.

Chapter 4 proposes to automatically discriminate between pathological and typical healthy speech by analyzing spectro-temporal subspaces of speech by applying a subspace-based discriminant analysis on the extracted subspaces.

Chapter 5 proposes two deep learning frameworks aiming at pathological speech detection. In the first approach, frame-level distance patterns between phonetically-balanced articulatory feature representations from healthy and test speakers are analyzed and classified using a neural network. Feature extraction, distance matrix computation, and classification are jointly optimized in an end-to-end framework. In the second approach, supervised representation learning frameworks with two auxiliary tasks are explored. To obtain a speaker identity-invariant representation, an adversarial auxiliary speaker identification task is used, while to obtain a discriminative representation for the speech pathology detection task, an auxiliary pathological speech classifier is used. Due to the challenges of adversarial training, a feature separation framework is also proposed where a speaker identity-invariant representation can

be obtained without using any adversarial training.

Chapter 6 proposes a short-time objective intelligibility measure to automatically assess the intelligibility of pathological speech by comparing the test speech representations to an intelligible reference model created from fully intelligible speech signals. To increase the applicability of the measure for different scenarios, the feasibility of creating the reference model from synthetic speech signals is also explored.

Chapter 7 proposes to use subspace-based analysis to develop a pathological speech intelligibility measure. Our proposed subspace-based intelligibility measure is based on a sub-Grassmannian subspace distance measure between subspaces spanning the fully-intelligible and pathological speech representations. Further in the same chapter, it is shown that the proposed measure is applicable to different scenarios and can also capture the effects of spectral modulation degradation that are important to the perceived speech intelligibility. In addition, two variants of the subspace-based intelligibility measures are explored by incorporating short-time temporal information.

Chapter 8 presents the joint experimental analyses regarding pathological speech detection and intelligibility assessment as a step toward developing an automated clinical tool.

Chapter 9 concludes the thesis along with suggesting directions for future research.

Publications based on this thesis work

Chapter 4:

- **Janbakhshi**, P., Kodrasi, I., and Bourlard, H. (2020b). Subspace-based learning for automatic dysarthric speech detection. *IEEE Signal Processing Letters*, 28(1):96–100

Chapter 5:

- **Janbakhshi**, P., Kodrasi, I., and Bourlard, H. (2021). Automatic dysarthric speech detection exploiting pairwise distance-based convolutional neural networks. In *Proc. 46th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 7328–7332, Virtual Conference
- **Janbakhshi**, P. and Kodrasi, I. (2021). Supervised speech representation learning for Parkinson's disease classification. In *Proc. 14th ITG Conference on speech communication*, pages 1–5, Virtual Conference

Chapter 6:

- **Janbakhshi**, P., Kodrasi, I., and Bourlard, H. (2019a). Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. In *Proc. 44th IEEE*

International Conference on Acoustics, Speech, and Signal Processing, pages 6405–6409, Brighton, UK

- **Janbakhshi**, P., Kodrasi, I., and Bourlard, H. (2020c). Synthetic speech references for automatic pathological speech intelligibility assessment. In *Proc. 45th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6099–6103, Virtual Conference

Chapter 7:

- **Janbakhshi**, P., Kodrasi, I., and Bourlard, H. (2019b). Spectral subspace analysis for automatic assessment of pathological speech intelligibility. In *Proc. 20th Annual Conference of the International Speech Communication Association*, pages 3038–3042, Graz, Austria

- **Janbakhshi**, P., Kodrasi, I., and Bourlard, H. (2020a). Automatic pathological speech intelligibility assessment exploiting subspace-based analyses. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28(1):1717–1728

2 Background on automatic pathological speech detection

In this chapter, we provide an overview of state-of-the-art automatic pathological speech detection approaches and establish the experimental frameworks and settings for the pathological speech detection approaches proposed in this thesis. We summarize several techniques proposed in the literature in Section 2.1. Then, we describe different databases, the required preprocessing steps, and validation strategies used to train and evaluate our proposed pathological speech detection approaches in Section 2.2. The evaluation metrics are summarized in Section 2.3. The details of different acoustic speech representations used to represent the speech signals for our proposed detection approaches are provided in Section 2.4. Finally, a summary of the chapter is presented in Section 2.5.

2.1 Literature overview

Research on automatic pathological speech detection over the last decades can be divided into two broad categories. One line of research is based on classical machine learning approaches where first acoustic features characterizing different impaired speech dimensions are handcrafted, and then classical classifiers using these handcrafted acoustic features are trained to discriminate between pathological and healthy speech. The used acoustic features can be either knowledge-driven features related to distorted speech dimensions or can be data-driven features resulting from modeling the short-time feature representations (at the utterance or speaker level) by different techniques (Hegde et al., 2019; Gómez-García et al., 2019). The other line of research focuses on pure data-driven deep learning approaches by seeking to exploit high-level abstract representations for automatic speech pathology detection. The feature design and the choice of data-driven approaches depend on the pathological speech task under study. Hence, in the following, we first introduce two popular speech tasks that have been commonly used for automatic pathological speech detection.

2.1.1 Speech tasks in pathological speech analysis

Most studies have focused on analyzing abnormal vocal fold function in pathologies such as laryngeal diseases (Dejonckere and Lebacq, 1996; Michaelis et al., 1998; A. Dibazar et al., 2002; Godino-Llorente and Gomez-Vilda, 2004; Dibazar et al., 2006; Arias-Londoño et al., 2010; Arjmandi and Pooyan, 2012; Ali et al., 2016; Travieso et al., 2017). For this reason, majority of the related studies are based on analyzing sustained vowels (sustained phonation) data. Sustained phonation tests in analyzing such pathologies are popular because they are independent of the language, speaking rate, and complicated articulatory behaviour (Parsa and Jamieson, 2001). These tests can elicit abnormal voice production symptoms which are also common distortions in dysarthria of speech, therefore sustained vowel analysis has been also used for detecting speech disorders such as PD (Little et al., 2009; Travieso et al., 2017; Almeida et al., 2019; Karan et al., 2020c, 2021). Although phonation analysis in the above state-of-the-art literature achieved acceptable performance, dysarthria affects several components of the speech production mechanism. Therefore, analyzing only sustained vowels might not be enough for characterizing dysarthric speech due to different diseases. On the other hand, analyzing connected speech (i.e., sentences and isolated words) can be expected to characterize many impaired dimensions of the overall speech production system in such disorders (Enderby, 2013; Godino-Llorente et al., 2017a).

Unlike sustained vowels, analysis of connected speech characterizing important dynamic aspects of vocal function faces several challenges. First, connected speech analyses are not as controlled as sustained phonation analyses due to many non-stationary vocal variations. Second, segmentation of voiced/unvoiced and silent regions of speech is often needed for such analysis. These segmentations should be performed with caution because accurate voicing detection can be difficult due to the inherent low quality of severe pathological speech (Duffy, 2000; Parsa and Jamieson, 2001; Little et al., 2009). Finally, inter-speaker variabilities such as language, accent, and habitual speaking rate are more likely to affect the performance of automatic systems based on connected speech analyses.

In this thesis, we focus on developing approaches applicable to connected speech. Hence, in the following section, an overview of state-of-the-art approaches using connected speech for assessing speech disorders is provided.

2.1.2 Machine learning-based approaches

As mentioned before, typical automatic pathological speech detection techniques are based on classical machine learning approaches operating on acoustic features which are handcrafted to reflect different impaired speech dimensions. In the following, we give an overview of commonly used acoustic features followed by classification techniques for pathological speech detection.

Feature design

Motivated by quantifying impacted phonation and voice quality degradation, many classical acoustic features have been designed. To characterize the perceptual roughness of voiced speech which is associated with irregularity or aperiodicity in the vocal cord vibration, fundamental frequency f_0 and short-term variations of the fundamental frequency (jitter) and of the cycle-to-cycle peak amplitude (shimmer) have been used. To characterize the perceptual breathiness of speech caused by glottal air leakage, features such as harmonics-to-noise ratio (HNR), normalized noise energy, and low-to-high energy ratio measures have also been used for automatic pathological speech detection (Dejonckere and Lebacq, 1996; Michaelis et al., 1998; Little et al., 2009; Tsanas et al., 2012; Bocklet et al., 2013; Wang et al., 2016; Orozco-Arroyave et al., 2016a; Gillespie et al., 2017; Berus et al., 2019; Karan et al., 2020a). For analyzing connected speech, extracting the above-mentioned features requires voice/unvoiced segmentation of the speech signal (Wang et al., 2016; Orozco-Arroyave et al., 2016a; Berus et al., 2019). Therefore, computing such features can be unreliable due to the difficulty in accurately estimating the pitch period in pathological speech.

In addition, to characterize impacted articulation, features classically used in speech recognition have been popular for pathological speech detection, i.e., Mel frequency Cepstral coefficients and linear prediction coefficients (A. Dibazar et al., 2002; Godino-Llorente and Gomez-Vilda, 2004; Orozco-Arroyave et al., 2014b, 2015b,a; Wang et al., 2016; Gillespie et al., 2017; Illa et al., 2018; Karan et al., 2020a).

To jointly quantify impacted phonation and articulation, the sparsity of speech characterized through the shape parameter of the distribution of speech spectral coefficients has been used for pathological speech detection (Kodrasi and Bourlard, 2019, 2020).

Aiming to capture as many impaired dimensions as possible, large-scale feature sets such as openSMILE have also been used in the literature for pathological speech detection (Bocklet et al., 2013; Wang et al., 2016; Vaiciukynas et al., 2017b; Norel et al., 2018; Berus et al., 2019). For example, openSMILE brute-forced acoustic feature set containing 6373 features resulting from the computation of statistics over low-level descriptor contours is introduced as a baseline feature set for 2013 Interspeech Computational Paralinguistics Challenge (ComParE) targeting many paralinguistic tasks, e.g., emotion and autism classification from speech (Eyben et al., 2013; Schuller et al., 2013).

Classification approaches

Automatic speech pathology detection is a binary classification problem. The classification approaches to detect pathological speech range from simple classifiers such as k-nearest neighbors (Arjmandi and Pooyan, 2012), Gaussian mixture models (GMMs) (A. Dibazar et al., 2002; Ali et al., 2016; Godino-Llorente et al., 2017b) to more complex ones such as artificial neural networks (ANNs) and hidden Markov models (HMMs) (A. Dibazar et al., 2002; Dibazar

et al., 2006; Arias-Londoño et al., 2010; Travieso et al., 2017; Illa et al., 2018; Berus et al., 2019). One of the most popular classifiers in the literature for this task is kernel support vector machines (SVMs) (Karan et al., 2020c; Arjmandi and Pooyan, 2012; Bocklet et al., 2013; Orozco-Arroyave et al., 2014b, 2016b; Wang et al., 2016; Travieso et al., 2017; Illa et al., 2018; Norel et al., 2018; Kodrasi and Bourlard, 2020).

Since most acoustic features are designed to parametrize short segments (i.e., frames) of speech (referred to as short-time features), depending on the type of the classifier, short-time feature representations need to be aggregated at the utterance or speaker level before feeding the input to the classifier. A common feature aggregation method is computing statistical functions such as the mean, standard deviation, skewness, and kurtosis of the short-time features across the utterance length to obtain a fixed-length feature vector for each utterance (Bocklet et al., 2013; Orozco-Arroyave et al., 2014b, 2016b; Wang et al., 2016; Vaiciukynas et al., 2017b; Norel et al., 2018; Berus et al., 2019). Even after such feature aggregation, a large number of features can remain, which increases the risk of over-fitting due to the scarcity of pathological speech training data. Therefore feature selection and feature dimensionality reduction methods are often used prior to training the classifiers, e.g., correlation filtering, linear discriminant analysis (LDA), and principal component analysis (PCA) (Little et al., 2009; Arjmandi and Pooyan, 2012; Tsanas et al., 2012; Wang et al., 2016; Norel et al., 2018; Berus et al., 2019). Furthermore, inspired by the speaker recognition techniques, short-time feature aggregation (with the possibility of being unified with the classification step) is also achieved by GMM-universal background models (GMM-UBM) and iVector-based modeling (García et al., 2017; Godino-Llorente et al., 2017b; Moro-Velázquez et al., 2018). GMM modeling followed by Fisher vector encoding is also exploited in Egas-López et al. (2019) for this task.

2.1.3 Deep learning-based approaches

As mentioned in the previous section, typical contributions for automatic pathological speech detection are based on handcrafting acoustic features. Such features may fail to adequately capture pathological speech characteristics. Further, since handcrafted features are based on clinicians' knowledge, they may also fail to characterize abstract but important acoustic cues present in pathological speech. As an alternative to using handcrafted acoustic features, high-level abstract representations of speech can be extracted using data-driven deep learning approaches (Vasquez et al., 2017; Vaiciukynas et al., 2017a; Cummins et al., 2018; An et al., 2018; Bhati et al., 2019; Mallela et al., 2020; Vasquez-Correa et al., 2020; Karan et al., 2020b). Although deep learning has improved performance over conventional machine learning methods in many speech applications, they have not yet had the expected dominating influence in the pathological speech assessment field. The latter can be explained by the relatively small size of pathological speech databases compared to the databases used for other speech tasks (Cummins et al., 2018). Therefore, the main challenge in successfully exploiting deep learning approaches in pathological speech assessment is alleviating overfitting issues associated with the typically limited training data that is available. To increase the number of training samples,

speech signals are split into short segments (e.g., 160 ms), each segment is labeled as healthy or pathological depending on the label of the complete signal, and convolutional neural networks (CNNs) are trained on these segments for pathological speech detection (Vasquez et al., 2017; Vaiciukynas et al., 2017a; An et al., 2018). A similar approach is also used in Mallela et al. (2020) where cascaded CNN and long short-term memory (LSTM) layers are exploited to classify the speech segments. In Bhati et al. (2019), LSTM Siamese networks are used for pathological speech detection, where networks with Siamese architectures are trained on pairs of input data with the same phonetic content. Pairwise training in such networks helps to extract features that are discriminative of pathological speech while being robust to non-relevant information in the limited resource scenario. However, since input data needs to have the same phonetic content, different LSTM networks need to be trained for different utterances.

In addition, unsupervised representation learning is also exploited to extract features from short (phonetically unmatched) segments of speech representations for pathological speech detection. Feature representations are first learned using CNN auto-encoders trained with a large amount of healthy speech data. These representations are then extracted for training a separate pathological speech classifier (Vasquez-Correa et al., 2020). A similar approach is also used in Karan et al. (2020b) where instead of typical auto-encoders, stacked auto-encoders are exploited. Using unsupervised representation learning approaches, there is no guarantee that the extracted representations are discriminative enough for pathology detection. In Korzekwa et al. (2019), a supervised representation learning framework is used where two encoders, i.e., an audio and a text encoder are exploited to generate representations that not only can reconstruct the input segments but also have discriminative information regarding pathological speech. However, such an approach requires text transcriptions for training, which can limit its applicability for scenarios where the speech transcription is not available.

2.2 Databases and protocols

Despite the quantity and quality of the state-of-the-art approaches for pathological speech detection, their performance evaluation is typically limited to one database, i.e., speech data with one language or impaired speech data resulted from one speech disorder, therefore, their generalization can be limited. In addition, evaluation protocols used in state-of-the-art approaches largely vary causing the difficulty of a fair comparison of different approaches; while each approach may have revealed promising conclusions based on their used private database or evaluation strategy. This can be explained by the lack of standard and open-access pathological speech databases and their corresponding established evaluation protocols.

In this thesis, for evaluation of pathological speech detection approaches and assessing their generalisability, we use two databases with different languages. By unifying our evaluation protocols and the pre-processed databases, we aim to achieve a fair evaluation of different pathological speech detection approaches. In the following, the data sets and protocols used

for the pathological speech detection task in this thesis are provided.

2.2.1 PC-GITA database

The PC-GITA database (Orozco-Arroyave et al., 2014a) consists of recordings from 50 PD patients (25 males, 25 females) and 50 healthy speakers (25 males, 25 females). All speakers are adults and Colombian Spanish native speakers. The recordings were captured in a soundproof booth. All of the patients were diagnosed by neurologists and they were recorded with the patients in ON-state, i.e. no more than 3 hours after the morning medication. None of the healthy controls had symptoms associated with any neurological diseases. Each speaker utters 24 isolated words, 6 sentences, 4 sentences with additional emphasis on particular words, and 1 text, with all utterances being recorded at a sampling frequency of 44.1 kHz.

Preprocessing

All recordings are down-sampled to 16 kHz, and speech-only segments are extracted using an energy-based voice activity detector (Boersma, 2002) for all recordings except for words. For recordings of words in this database, speech-only segments were already manually extracted. Concatenating speech-only segments from all recordings for each speaker yields an average of 61.07 seconds long speech signal for the healthy speakers and an average of 59.93 seconds long speech signal for the patients.

Performance evaluation

To evaluate the performance of pathological speech detection approaches on this database, we use a stratified speaker-independent 10-fold cross-validation ensuring that each fold has the same number of healthy and pathological speakers and that there is no overlap between speakers across folds used for training and evaluation. However, in one of our approaches, i.e., the temporal subspace-based approach in Chapter 4, several healthy speakers should be used as references required for the time-alignment step. For a fair evaluation, we exclude the reference speakers from our final pathological speech detection evaluation. After excluding 5 randomly selected reference healthy speakers, we also exclude 5 randomly selected patients to keep the balance between the number of speakers in the two groups. To evaluate this approach, considering overall 45 PD patients (22 males, 23 females) and 45 healthy speakers (22 males, 23 females), we use a stratified speaker-independent 9-fold cross-validation framework.

2.2.2 MoSpeeDi database

The MoSpeeDi database consists of recordings from French-speaking adults including 20 PD and ALS patients (14 males, 6 females) and 30 healthy speakers (11 males, 20 females) from Geneva University Hospitals and University of Geneva. Each speaker utters 54 pseudo-words

and 8 sentences based on the MonPaGe speech screening protocol (Fougeron et al., 2018), with all utterances being recorded at a sampling frequency of 44.1 kHz. This database is collected as a part of the interdisciplinary SNF Sinergia project on motor speech disorders.

Preprocessing

After downsampling all recordings to 16 kHz, speech-only segments are extracted from the recordings using an energy-based voice activity detector (Boersma, 2002). Concatenating speech-only segments from all recordings for each speaker yields an average of 103.83 seconds long speech signal for the healthy speakers and an average of 115.28 seconds long speech signal for the patients.

Performance evaluation

To ensure a balanced number of healthy and pathological speakers, we exclude 10 healthy speakers (9 females, 1 male) before evaluating the performance of our pathological speech detection approaches on this database. Hence, considering 20 patients (14 males, 6 females) and 20 healthy speakers (10 males, 10 females), we use a stratified speaker-independent 5-fold cross-validation framework ensuring that each fold has the same number of healthy and pathological speakers and that there is no overlap between speakers across folds used for training and evaluation. As described in Section 2.2.1, one of our proposed pathological speech detection approaches, i.e., temporal subspace-based approach in Chapter 4, requires healthy reference speakers. The data from the excluded 10 healthy speakers are used as references required for the time-alignment step in this approach.

2.3 Evaluation metrics

The performance of the pathological speech detection task as a binary classification problem is evaluated by two metrics: i) classification accuracy, i.e., the percentage of correctly predicted (pathological vs. healthy) speakers. Due to the class balance for the above-mentioned databases, accuracy measure is a suitable non-biased metric for our evaluation, and ii) area under ROC curve (AUC) measuring the classification performance at various threshold settings.

2.4 Speech representations

In pathological speech detection approaches in this thesis, different speech representations are exploited. The usage of some representations such as short-time Fourier transform, Mel-scale representation, and Mel frequency Cepstral coefficients are motivated by their success in state-of-the-art approaches. Therefore, in line with state-of-the-art approaches, we further evaluate the efficacy of such representations in our frameworks. In addition, depending on our

used frameworks, we also propose to use other representations that have not been explored for such a task before, i.e., one-third octave band and articulatory posterior representation. This section gives an overview of the speech representations used in our pathological speech detection approaches.

2.4.1 Short-time Fourier transform representation

Short-time Fourier transform (STFT) (Allen and Rabiner, 1977) is the most commonly used time-frequency (TF) representation in speech processing. To obtain the STFT representation of a discrete-time signal $s(n)$, the speech signal is first segmented into fixed-length (usually overlapping) frames. After weighting each frame by an analysis window $w_{\text{stft}}(n)$, the discrete Fourier transform (DFT) is applied to each frame yielding the complex time-frequency coefficients $S_{\text{stft}}(f,m)$, i.e.,

$$S_{\text{stft}}(f,m) = \sum_{n=0}^{N-1} w_{\text{stft}}(n)s(mR+n)e^{-\frac{i2\pi nf}{F}}, \;\; f \in \{0,...,F-1\}, \tag{2.1}$$

with f being the index of the frequency bin, m being the time frame index, F being the total number of frequency bins, N being the frame size (length of the window), R being the frame shift, and i being the imaginary unit, i.e., $i^2 = -1$. The complex STFT coefficients can be expressed as $S_{\text{stft}}(f,m) = |S_{\text{stft}}(f,m)|e^{i\theta(f,m)}$, with $|S_{\text{stft}}(f,m)|$ and $\theta(f,m)$ denoting the magnitude and phase of the TF units. For this thesis, we only use the magnitude of TF units. However, as we have shown in Janbakhshi and Kodrasi (2022), also the phase information can be exploited for pathological speech detection.

2.4.2 One-third octave band representation

To obtain a signal representation resembling the transform properties of the human auditory system, the logarithm of one-third octave band representation is used. To obtain the one-third octave band representations, the signals are first transformed to the TF domain using the STFT and then one-third octave band analysis is applied to the STFT representation (Elliott and Theunissen, 2009; Jensen and Taal, 2016), i.e.,

$$S_{\text{oct}}(j,m) = \log_{10}\sqrt{\sum_{f\in \text{CB}_j} |S_{\text{stft}}(f,m)|^2}, \tag{2.2}$$

where j denotes the one-third octave band index, CB_j denotes the indices of STFT coefficients corresponding to the j^{th} one-third octave band, and $j \in \{0,..,J-1\}$ where J is the total number of octave bands.

2.4.3 Mel-scale representation

Alternatively to the STFT representation, we also use log Mel-scale representation where nonlinear frequency scaling motivated by the human auditory system is applied to the STFT representation using a set of overlapping triangular filters. The central frequencies of the filters are equally spaced in the Mel-frequency domain. A given frequency in the Mel scale corresponds to the frequency in Hz based on the relation $\nu(\text{Hz}) = 700e^{\nu(\text{Mel})/1127-1}$ (Makhoul and Cosell, 1976). Denoting the k^{th} Mel-filter with Λ_k, $k \in \{0,...,K\}$ where K is the total number of filters, the Mel-scale representation is computed as

$$S_{\text{Mel}}(k,m) = \log_{10} \sum_f |S_{\text{stft}}(f,m)|^2 \Lambda_k(f). \tag{2.3}$$

2.4.4 Mel frequency cepstral coeffcients

Mel frequency Cepstral coefficients (MFCCs) capturing the vocal tract characteristics are widely used in speech processing applications. MFCCs are extracted by computing the discrete cosine transform of the logarithm of the (previously described) Mel-scale representation in (2.3) and selecting the first K computed coefficients (Davis and Mermelstein, 1980).

2.4.5 Articulatory posterior representation

Distinct speech sounds are produced by different articulatory properties, i.e., by different vocal tract configurations of the lips, jaw, tongue, pharynx, and palate altering the resonances of the vocal tract (Pasley et al., 2015). Such articulatory properties of the vocal tract to produce distinct sounds can be characterized by articulatory features. Articulatory features can be quantified by learning the mappings between linguistic sub-word units of any language such as phonemes to articulatory properties such as the place of constriction, the height of the tongue, roundedness of the lips, etc. Articulatory representations in this thesis are extracted as in Dubagunta and Magimai-Doss (2019), where frame-level posteriors of four articulatory categories are computed, i.e., manner of articulation (e.g., degree of constriction), place of constriction, the height of the tongue, and vowel. Articulatory posteriors (AP) for each category are estimated using CNNs trained on healthy speech data from the English AMI corpus (Carletta et al., 2005) based on acoustic phoneme-to-articulatory feature mappings (Rasipuram and Magimai.-Doss, 2016). To obtain the final AP representation per time frame, all extracted APs for each category are concatenated. For details on the training procedure for AP feature extraction, the reader is referred to Dubagunta and Magimai-Doss (2019).

2.5 Summary

In this chapter, we have discussed several research areas on state-of-the-art automatic pathological speech detection approaches. The majority of the approaches are based on classical machine learning where handcrafted acoustic features are first extracted and used to train classical classifiers to achieve pathological and typical speech discrimination. For many of the handcrafted acoustic features related to vocal source, voiced speech segmentation methods are required which can fail due to the low quality of pathological speech. More recently, there has been a growing interest in leveraging deep learning-based approaches for pathological speech assessment, where neural networks have been exploited to extract high-level and abstract speech representations. However, due to challenges in successfully applying deep learning approaches to pathological speech assessment, fewer contributions have been made when compared to approaches based on classical machine learning. In the majority of deep learning-based approaches for speech pathology detection, the neural networks are not explicitly trained to extract robust features. In this chapter, we have also introduced databases and evaluation metrics used to evaluate the proposed detection approaches in the remainder of this thesis. Details on different acoustic speech representations used in this thesis have also been provided.

3 Background on automatic pathological speech intelligibility assessment

In this chapter, we provide an overview of state-of-the-art automatic pathological intelligibility techniques and establish the experimental framework and settings for the intelligibility assessment techniques proposed in this thesis. We summarize several intelligibility measures proposed in the literature in Section 3.1. Then, different databases, the required preprocessing steps, and validation strategies used to evaluate our proposed intelligibility measures are presented in Section 3.2. The evaluation metrics are described in Section 3.3 and the information about the used acoustic speech representation for the proposed intelligibility measures is provided in Section 3.4. Finally a summary of the chapter is presented in Section 3.5.

3.1 Literature overview

In the past decade, several approaches for the automatic assessment of pathological speech intelligibility have been proposed. Such approaches aim to develop an objective intelligibility measure correlated with the subjective intelligibility scores, i.e., the percentage of words correctly understood by listeners. These approaches can be broadly categorized into blind approaches (Paja and Falk, 2012; Hummel et al., 2011; Falk et al., 2012; Martínez et al., 2013; Kim et al., 2014; Haderlein et al., 2017; Fletcher et al., 2017) and non-blind approaches (Haderlein et al., 2004; Middag et al., 2008, 2009; Maier et al., 2009; Nuffelen et al., 2009b; Middag et al., 2010; Bocklet et al., 2012; Martínez et al., 2015; Imed et al., 2017; Kalita et al., 2018).

Blind approaches refer to approaches that do not exploit any knowledge about healthy (i.e., intelligible) speech and assess pathological speech intelligibility by extracting acoustic features that are believed to be correlated with intelligibility. In Hummel et al. (2011); Falk et al. (2012); Paja and Falk (2012); Haderlein et al. (2017); Fletcher et al. (2017), individual acoustic features such as jitter, shimmer, fundamental frequency, formant frequencies, voiced frames percentage, or low-to-high modulation energy ratio (LHMR) are directly used to assess pathological speech intelligibility. In Hummel et al. (2011); Paja and Falk (2012); Falk et al. (2012); Kim et al. (2014); Fletcher et al. (2017), multiple acoustic features are handcrafted and combined through feature selection/reduction methods and then regression models are

trained to estimate speech intelligibility. It should be noted that many of the handcrafted acoustic features are similar to the ones used in classical machine learning-based approaches for speech pathology detection discussed in Setion 2.1.2. However, instead of training a binary classifier to discriminate between healthy and pathological speech, intelligibility assessment deals with a regression problem to predict the speech intelligibility of pathological speakers. Although several proposed measures have been shown to be correlated with subjective intelligibility scores, rigorous validation strategies have not always been followed. For example, a fair leave-one-subject-out paradigm or a separate test and train set have not been reported for feature selection or regression training (Hummel et al., 2011; Paja and Falk, 2012; Falk et al., 2012).

Non-blind approaches rely on intelligible speech recordings from healthy speakers to estimate pathological speech intelligibility. In these approaches, healthy speech recordings are exploited in different manners. In Bocklet et al. (2012), a speaker-independent GMM is trained on healthy speech to create an intelligible reference model. By adapting the parameters of this reference model, a GMM-based supervector is created to represent the pathological speech signal. The intelligibility score is then obtained by training a regression model on the GMM-based supervector. A very similar approach is followed in Martínez et al. (2013); Martínez et al. (2015); Imed et al. (2017); Kalita et al. (2018), with the difference consisting in using an iVector or Gaussian posteriogram representation instead of a GMM-based supervector. In other non-blind approaches, pathological speech intelligibility is evaluated by training regression models on features produced by automatic speech recognition (ASR) systems, automatic speech alignment (ASA) systems, or phonological feature (PLF) extractor systems (Haderlein et al., 2004; Schuster et al., 2005; Windrich et al., 2008; Middag et al., 2008; Maier et al., 2009; Middag et al., 2009; Nuffelen et al., 2009b; Middag et al., 2010; Kim et al., 2015; Dimauro et al., 2017). Commonly used features from such systems are the word error rate (WER), log-likelihood ratio, phoneme posteriors, and phonological features. These systems are typically trained using a large number of transcribed/segmented healthy speech recordings (Haderlein et al., 2004; Schuster et al., 2005; Windrich et al., 2008; Middag et al., 2008, 2009; Maier et al., 2009; Nuffelen et al., 2009b; Middag et al., 2010; Kim et al., 2015; Dimauro et al., 2017).

Although promising results have been shown using the above-mentioned approaches, several drawbacks arise when using them in practical scenarios. Most approaches require a large number of features for intelligibility prediction, increasing as a result the risk of over-fitting and limiting the performance in unseen data due to very limited pathological training data for intelligibility assessment. In addition, non-blind approaches are typically complex and require a large number of healthy speech recordings for training, which might be infeasible for low-resource languages. Finally, non-blind approaches relying on ASR, ASA, and PLF systems require transcriptions of healthy and/or of pathological speech signals, which can be a time- and resource-consuming task.

3.2 Databases and protocols

Although many state-of-the-art pathological speech intelligibility measures have been proposed, their performance evaluation is typically limited to one database, i.e., speech data with one language or impaired speech data resulted from one speech disorder, therefore, their generalization can be limited. In addition, evaluation protocols used for state-of-the-art intelligibility measures are very diverse which results in the difficulty of conducting a fair comparison between different measures. Similarly to the speech pathology detection task described in the previous chapter, such diversity can be explained by the lack of standard and open-access pathological speech databases for intelligibility assessment and their corresponding established intelligibility evaluation protocols.

In this thesis, for evaluation of pathological speech intelligibility measures and assessing their generalisability, we use two databases with different languages and disorders. To evaluate the proposed intelligibility measures, the ground truth for speech intelligibility, i.e., subjective intelligibility scores, are required. Therefore, only databases with available subjective intelligibility scores computed through subjective listening paradigms can be used. Such subjective intelligibility scores are not available for the previously mentioned databases in Section 2.2 that are considered for pathological speech detection task in Chapter 2. Hence, different databases are considered for pathological speech intelligibility assessment task.

In this thesis, by unifying our evaluation protocols and the pre-processed databases, we aim to achieve a fair evaluation of different approaches for pathological speech intelligibility assessment. In the following, the data sets and protocols used for the pathological speech intelligibility assessment task in this thesis are provided.

3.2.1 Universal access speech (UA-Speech) database

The UA-Speech database (Kim et al., 2008) includes recordings of 15 adult English-speaking dysarthric patients (11 males, 4 females) diagnosed with CP and of 13 adult healthy speakers (9 males, 4 females). Each speaker read 763 isolated words, with 155 of the words uttered three times and referred to as common words (CW). The remaining 298 words were uttered only once and are referred to as uncommon words (UW). A 7-channel microphone array is used for recording the speakers at a sampling rate of 16 kHz. The subjective intelligibility scores of patients are computed by performing a subjective listening test using 5 naive listeners for each patient. Listeners provided orthographic transcriptions of speech utterances, and based on the mean percentage of the correct transcribed responses across the listeners, the subjective intelligibility score of each patient is obtained. The subjective intelligibility scores of patients range from 2% to 95%.

Preprocessing

For evaluating intelligibility assessment approaches in this thesis, we consider the recordings of the (arbitrarily selected) 5th channel from the UA-Speech database. To extract speech-only segments, an energy-based voice activity detection (Boersma, 2002) is applied to the speech recordings. Concatenating speech-only segments from all recordings for each speaker yields an average of 6199.8 seconds long speech signal for the healthy speakers and an average of 17084.6 seconds long speech signal for the patients.

Validation strategy

For the UA-Speech database two scenarios are considered.

Phonetically-balanced scenario. In this scenario all speakers (healthy and pathological) utter exactly the same words. All 763 available words are considered for this database. The intelligibility score is calculated for each word, and the final intelligibility score for each patient is computed as the mean intelligibility score across all words.

Phonetically-unbalanced scenario. In this scenario, we assume that all speakers (healthy and pathological) utter different sets of word utterances. For word-level intelligibility assessment, the intelligibility score is calculated for each word uttered by each patient, and the final intelligibility score is computed as the mean intelligibility score across all available words for that patient. If the word-level intelligibility assessment is not possible (depending on the proposed measures, cf. Chapter 7), different sets of words are concatenated to create longer utterances for each speaker, and a single intelligibility score is estimated for each patient.

Since the UA-Speech database contains a large number of words that can be combined in different ways for different speakers, phonetically-unbalanced analyses are done on the UA-Speech database.

3.2.2 Dutch corpus of pathological and normal speech (COPAS)

From the COPAS database (Nuffelen et al., 2009a), we consider recordings of 16 adult Dutch-speaking HI patients with a speech disorder (6 males, 10 females) and of 22 adult healthy speakers (11 males, 11 females). For each speaker, recordings of 10 sentences sampled at 16 kHz are used. The subjective intelligibility scores of the speakers are computed based on the Dutch intelligibility assessment (DIA) test where a listener identifies missing phonemes in word templates of 50 monosyllabic words uttered by each speaker. Subjective intelligibility is then defined as the percentage of correctly identified phonemes for each speaker. The subjective intelligibility scores of patients range from 53% to 98%.

Preprocessing

Individual words are extracted from all sentences using forced alignment from an ASR system followed by manual corrections, resulting in 47 available words for each speaker. Concatenating all available words for each speaker yields an average of 22.9 seconds long speech signal for the healthy speakers and an average of 31.1 seconds long speech signal for the patients.

Validation strategy

For this database, only phonetically-balanced analysis is considered since unlike the UA-Speech database, it does not contain enough words to combine in different ways for creating phonetically-unbalanced scenarios. Given that all speakers (healthy and pathological) utter exactly the same 47 words, the intelligibility score is calculated for each word, and the final intelligibility score for each patient is computed as the mean intelligibility score across all words.

3.3 Evaluation metrics

To evaluate the performance of the automatic pathological intelligibility measures, the Pearson correlation coefficient (R) and the Spearman rank correlation coefficient (R_S) between the automatically estimated intelligibility and the subjective intelligibility scores of the patients are computed. In addition, the statistical significance of these correlation values is also assessed.

To evaluate the statistical significance, the critical values of R and R_S, denoted by R_c and R_{Sc}, respectively, are computed using a significance level $\alpha = 0.05$ and taking into account the number of patients in each database (Zwillinger and Kokoska, 2000a,b). The obtained critical values are presented in Table 3.1. The correlation values obtained for the different intelligibility measures are considered to be statistically significant if $|R| \geq |R_c|$ and $|R_S| \geq |R_{Sc}|$.

Table 3.1 – Critical values for the Pearson and Spearman correlation coefficients obtained using $\alpha = 0.05$ (Zwillinger and Kokoska, 2000a,b). The number of pairs of scores is considered to be the number of patients in each database. The correlation values obtained for any intelligibility measure are considered to be statistically significant if $|R| \geq |R_c|$ and $|R_S| \geq |R_{Sc}|$.

15 English CP patients		16 Dutch HI patients	
R_c	R_{Sc}	R_c	R_{Sc}
−0.441	−0.443	−0.426	−0.443

3.4 Speech representations

For our proposed intelligibility measures in this thesis, we consider a simple perceptually relevant acoustic speech representation, i.e., one-third octave band representation. The applicability of this speech representation has been established for objective speech intelligibility assessment in speech enhancement applications. One-third octave band representation of speech is described in the previous chapter (cf. Section 2.4.2).

3.5 Summary

In this chapter, we have discussed several state-of-the-art techniques for assessing pathological speech intelligibility which we believe to represent the most significant contributions in the field. Automatic pathological speech intelligibility assessment approaches can be broadly categorized into blind approaches and non-blind approaches. In blind approaches which do not require any healthy (intelligible) speech signals, several handcrafted acoustic features are extracted from pathological speech and are then analyzed to derive an intelligibility prediction. Non-blind approaches on the other hand encompass a wide range of approaches where healthy reference signals are exploited in different manners to extract features to be analyzed for intelligibility assessment. The performance of many of the state-of-the-art approaches can be limited since they are based on extracting a large number of acoustic features while using very limited pathological training data. In addition, the complexity of many non-blind approaches, their requirement to have access to a large number of healthy speech recordings, and/or to speech transcriptions limit their application for low-resource languages. Furthermore, in this chapter, we have also presented databases and evaluation metrics used to evaluate our proposed intelligibility measures in the remainder of the thesis.

4 Subspace-based learning for automatic pathological speech detection

In this chapter, we present our proposed approach to automatically discriminate between pathological and healthy speech based on analyzing speech spectral and temporal subspaces. The applicability and generalisability of the proposed subspace-based approach in this chapter are experimentally investigated across databases and also compared to using an SVM with state-of-the-art features.

4.1 Introduction

In Chapter 2 we introduced classical machine learning-based approaches for automatic pathological speech detection. In these approaches, short-time acoustic features are first handcrafted to reflect impaired speech dimensions related to the vocal source and tract. These features are then used to train a classifier for pathological speech detection. As mentioned in Chapter 2, due to the low quality of pathological speech, the pitch estimation or voiced speech segmentation methods that are required for features related to vocal source (e.g., f_0, jitter, shimmer or HNR) can fail (Parsa and Jamieson, 2001; Orozco-Arroyave et al., 2012). Furthermore, all extracted short-time (i.e., frame-level) features need to be aggregated at the utterance or speaker level to obtain fixed length representations before being fed to typical classifiers. Among common aggregation methods are statistical parametrization using statistical functionals and GMM-based modeling (cf. Section 2.1.2). In such feature aggregation, information regarding temporal patterns of short-time features is not captured although it can have important cues for pathological speech detection.

In this chapter we propose a learning method where spectral or temporal information can be separately modeled and exploited for pathological speech detection using a minimal number of parameters. Because of atypical changes in spectro-temporal fluctuations associated with imprecise and reduced articulatory movements in many speech disorders (e.g., dysarthria), the dominant spectro-temporal patterns of healthy and pathological speech can be expected to differ (Rosen et al., 2006). Therefore, we propose to extract spectro-temporal subspaces spanning the dominant spectro-temporal patterns of speech and use them as acoustic features

for automatic pathological speech detection. Spectro-temporal subspaces can be directly extracted from continuous speech without requiring voiced speech segmentation. Further, a subspace-based representation can be robust to unstructured random noise and can show better generalization performance without requiring a large amount of training data (Ruiping Wang et al., 2008; Chen et al., 2013; Mishra et al., 2019). Since we exploit the structural information embedded in the subspace models to discriminate between healthy and pathological speech, subspace-based learning yields a minimal number of parameters for training. To the best of our knowledge, a subspace-based learning framework for pathological speech detection has never been considered in the literature. Furthermore, while spectral subspaces are the typical choice for speech subspace analyses in many applications, temporal subspace analysis has never been explored. In Kacha et al. (2020), it has been experimentally shown that the mean of the first and second dominant spectral basis vector of healthy and pathological speech differ. However, no techniques aiming at automatic pathological speech detection using these spectral subspaces have been proposed.

The rest of the chapter is organized as follows. Section 4.2.1 describes the construction of spectro-temporal subspaces by extracting dominant basis vectors spanning the column (i.e., spectral) and row (i.e., temporal) space of the TF representation using singular value decomposition (SVD). The same section also provides a solution for unaligned utterances prior to constructing the temporal subspaces. Section 4.2.2 provides details on Grassmann discriminant analysis (GDA) which is a subspace-based discriminant analysis technique to automatically discriminate between pathological and healthy speakers with each speaker represented by subspaces. Experimental results are presented in Section 4.3, where it is shown that compared to spectral subspaces, temporal subspaces are more powerful discriminators for pathological speech detection. Section 4.5 presents a summary of the chapter.

4.2 Subspace-based pathological speech detection

As depicted in the schematic representation in Fig. 4.1, the proposed subspace-based pathological speech detection approach consists of computing spectro-temporal subspaces and applying subspace-based discriminant analysis using GDA. In the remainder of this section, the computational details of the proposed approach are presented.

4.2.1 Computing spectro-temporal subspaces

To obtain an acoustic feature representation, all speech signals are first transformed to the TF domain. Although any user-defined TF representations can be used, here we use either logarithm of the one-third octave band spectrum (cf. Section 2.4.2) or MFCC representations (cf. Section 2.4.4) that result into a low number of octave bands or MFCC coefficients. Let $\mathbf{S}_m$ denote the $(J \times N_m)$–dimensional TF representation of an utterance from speaker m, with J being the total number of frequency (e.g., one-third octave or MFCC) bands, N_m being the total number of time frames, and $J \ll N_m$. As the rank$(\mathbf{S}_m) = J$, selecting TF representations with

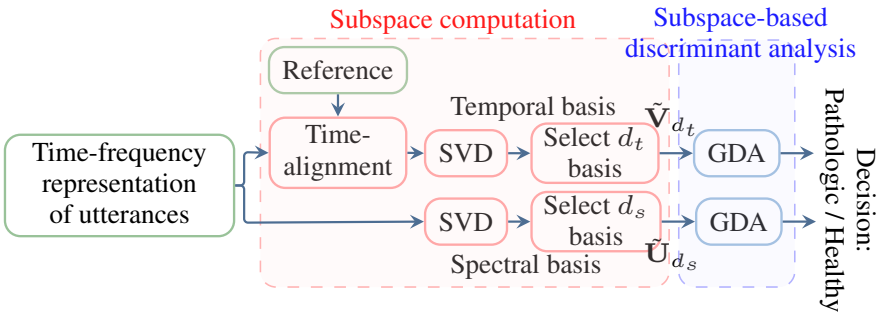


Figure 4.1 – Block diagram of the proposed subspace-based approach for pathological speech detection. The dominant spectro-temporal patterns of speech are characterized by subspaces spanning dominant spectral and temporal basis vectors of TF representations, where basis vectors are obtained using SVD. Considering subspaces as acoustic features, a subspace-based discriminant analysis, i.e., GDA, is used to automatically discriminate between pathological and healthy speakers.

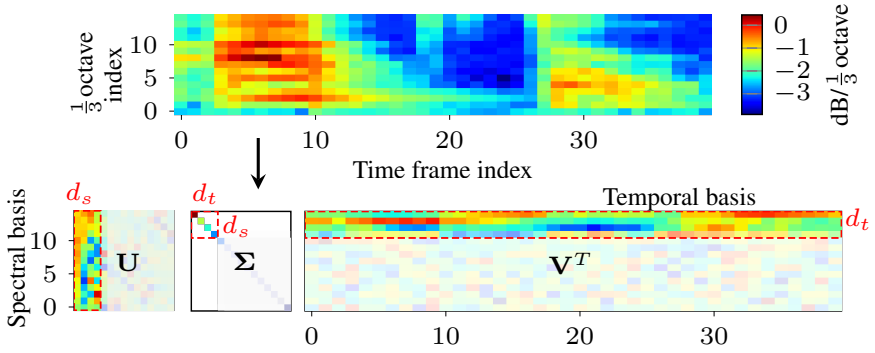


Figure 4.2 – Illustration of using SVD for obtaining spectral and temporal basis vectors spanning the spectral and temporal dimension of the TF representation of an utterance.

high values for J (e.g., in STFT representations) will result in higher computations required for tuning the number of spectral and temporal basis vectors as will be explained in the following subsections. While several techniques can be used to compute spectro-temporal basis vectors, in this chapter we propose to use SVD which provides an analytical solution and results in a high performance for our application. A schematic representation of applying the SVD to a sample utterance representation to obtain spectral and temporal basis vectors is depicted in Fig. 4.2.

The SVD of $\mathbf{S}_m$ is defined as

$$\mathbf{S}_m = \mathbf{U}\Sigma\mathbf{V}^T, \tag{4.1}$$

with $\mathbf{U}$ being the $(J \times J)$–dimensional orthonormal matrix of left singular vectors, Σ being the $(J \times J)$–dimensional diagonal matrix of singular values assumed to be sorted in descending

order, and $\mathbf{V}$ being the $(N_m \times J)$–dimensional orthonormal matrix of right singular vectors. Columns of $\mathbf{U}$ span the column space of $\mathbf{S}_m$, i.e., spectral space, and rows of $\mathbf{V}^T$ span the row space of $\mathbf{S}_m$, i.e., temporal space (Van Der Veen et al., 1993). Hence, in the following, columns of $\mathbf{U}$ and $\mathbf{V}$ will be referred to as spectral and temporal basis vectors, respectively [1]. It should be noted that for comparing spectral basis vectors across different speakers, no constraints on the speech phonetic content are required (on the condition that speech utterances are long enough to establish an average spectral pattern). However, the computation of temporal basis vectors highly depends on the phonetic content of the speech representations. Hence, comparing temporal basis vectors across different speakers requires the availability of the speech representations with the same phonetic content from all speakers.

Spectral subspaces

To construct the spectral subspace for speaker m, $\mathbf{S}_m$ is mean-centered in each frequency band prior to computing the SVD in (4.1). The $(J \times d_s)$–dimensional matrix $\tilde{\mathbf{U}}_{d_s}$ of dominant spectral basis vectors spanning the spectral subspace is then constructed from the first d_s spectral basis vectors in $\mathbf{U}$, where $d_s < J$ since $\text{rank}(\mathbf{S}_m) = J$. The parameter d_s can be automatically computed based on nested cross-validation (cf. Section 4.3.2).

It has been theoretically proven that computing spectral basis vectors using the SVD without mean-centering the representations biases the first spectral basis vector to the direction of mean spectral vector across time (also known as the long-term average spectrum (LTAS) in the speech community) rather than to the direction with maximal variability of spectral information (Cadima and Jolliffe, 2009; Alexandris et al., 2017). Furthermore, in Kacha et al. (2020), this phenomenon is experimentally confirmed for voiced segments of speech, i.e., the first principal component (PC) of the spectrogram is shown to be highly correlated with LTAS (Kacha et al., 2020). Although in Kacha et al. (2020) a group difference in the average of the first PC of the spectrograms (i.e., LTAS) computed from control speakers and PD patients is observed, in our application removing the bias of LTAS is important for achieving a good detection accuracy. As mentioned before, spectral mean is one of the statistical functionals commonly applied on TF features for feature aggregation. However, through initial analysis on our considered databases, we confirmed that using only the spectral mean is not discriminative enough for pathology detection. The difference in our finding and in Kacha et al. (2020) might be due to the fact that we consider all speech segments while in Kacha et al. (2020) only voiced frames are analyzed.

[1] It should be noted that spectral basis vectors given by (4.1) can be equivalently computed by PCA. PCA is commonly used for feature dimensionality reduction by projecting each feature vector (i.e., representing each speaker) onto the space spanned by the basis vectors of the feature matrix (consisting of features from all speakers). However, here we represent each speaker by a subspace spanned by basis vectors as the basic elements of our subspace-based learning method. In other words, here we are dealing with classifying subspaces represented by "a set" of vectors rather than (more conventionally) classifying feature vectors.

Temporal subspaces

As pointed out before, for computing and comparing temporal basis vectors, utterances from speakers must have the same phonetic content. The dominant temporal basis vectors in $\mathbf{V}$ from (4.1) can be used to construct the temporal subspace from $\mathbf{S}_m$. However, temporal basis vectors obtained from different speakers cannot be directly compared to each other because of unaligned TF representations (due to different speakers and speaking rates). Therefore, prior to computing the temporal basis vectors, we propose to time-align all TF representations using dynamic time warping (DTW) (Rabiner and Juang, 1993). Following a similar procedure as in (Kodrasi and Bourlard, 2020) for time-alignment, utterances $\mathbf{S}_m, m = 1, \ldots, M$, from all M available speakers are individually time-aligned to the $(J \times N_r)$–dimensional representation $\mathbf{S}_r$ of the same utterance from an (arbitrarily selected) healthy reference speaker r. For each time frame i in $\mathbf{S}_r$, with $i \in 1, ..., N_r$, all time frames in $\mathbf{S}_m$ that are mapped to it by DTW are extracted and averaged to create the corresponding time frame i in the time-aligned representation $\hat{\mathbf{S}}_m$. By repeating this procedure for all available $\mathbf{S}_m$, $m \neq r$, the utterance representations of all speakers are time-aligned. The dimension of the time-aligned representations $\hat{\mathbf{S}}_m$ is $J \times N_r$, i.e., it is dictated by the dimension of the reference representation $\mathbf{S}_r$.

To construct the temporal subspace for speaker m, the SVD is applied to the time-aligned representation as in (4.1), i.e.,

$$\hat{\mathbf{S}}_m = \hat{\mathbf{U}}\hat{\Sigma}\hat{\mathbf{V}}^T, \tag{4.2}$$

with $\hat{\mathbf{U}}$ being a $(J \times J)$–dimensional orthonormal matrix of spectral basis vectors, $\hat{\Sigma}$ being the $(J \times J)$–dimensional diagonal matrix of singular values assumed to be sorted in descending order, and $\hat{\mathbf{V}}$ being the $(N_r \times J)$–dimensional orthonormal matrix of temporal basis vectors. The time-aligned representations are mean-centered in each time frame prior to computing the SVD. The $(N_r \times d_t)$–dimensional matrix of dominant temporal basis vectors $\tilde{\mathbf{V}}_{d_t}$ spanning the temporal subspace is then constructed from the first d_t temporal basis vectors in $\hat{\mathbf{V}}$, where $d_t < J$ since $\text{rank}(\hat{\mathbf{S}}_m) = J$. The parameter d_t can be automatically computed based on nested cross-validation (cf. Section 4.3.2).

It should be noted that the computation of temporal subspaces relies on being able to accurately time-align representations. Based on our informal analyses, using DTW yields a very good alignment performance for our application.

4.2.2 Subspace-based discriminant analysis

Unlike typically used features that lie in a Euclidean space, subspaces lie in a non-Euclidean space called the Grassmann manifold. Hence, we propose to perform the classification for automatic pathological speech detection on this manifold using GDA (Hamm and Lee, 2008). GDA, which has shown promising results for image classification tasks, applies kernel linear discriminant analysis (LDA) using a Grassmann kernel respecting the geometry of subspaces on the manifold. The Grassmann manifold is first mapped into a high-dimensional Hilbert

space $\mathscr{H}$ which obeys the Euclidean geometry. This embedded manifold is then mapped into a lower-dimensional and more discriminative Euclidean space under the Fisher LDA criteria. Finally, the dimensionality-reduced data can be classified through classical classifiers such as LDA or k-nearest neighbors (Hamm and Lee, 2008).

For pathological speech detection, we are dealing with a two-class (healthy vs. pathological) classification problem where each class c, $c \in \{1,2\}$, has M_c training samples (speakers). Let $\mathbf{Y}_q$ denote the orthonormal matrix representing the (spectral or temporal) subspace associated with the training sample q. Further, let Φ denote the function mapping subspaces to the Hilbert space $\mathscr{H}$. Finding the discriminant Fisher directions $\mathbf{w}$ in $\mathscr{H}$ requires maximizing

$$J = \frac{\mathbf{w}^T \mathbf{S}_b^\phi \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w^\phi \mathbf{w}}, \tag{4.3}$$

with

$$\mathbf{S}_b^\phi = \frac{1}{M} \sum_{c=1}^{2} M_c (\mathbf{m}_c^\phi - \mathbf{m}^\phi)(\mathbf{m}_c^\phi - \mathbf{m}^\phi)^T, \tag{4.4}$$

$$\mathbf{S}_w^\phi = \frac{1}{M} \sum_{c=1}^{2} \sum_{\mathbf{Y}_q \in c} (\mathbf{Y}_q^\phi - \mathbf{m}_c^\phi)(\mathbf{Y}_q^\phi - \mathbf{m}_c^\phi)^T, \tag{4.5}$$

where $M = M_1 + M_2$, $\mathbf{m}_c^\phi$ denotes the mean of the mapped training samples from class c, $\mathbf{m}^\phi$ denotes the mean of all mapped training samples, and $\mathbf{Y}_q^\phi$ denotes the mapped training sample $\mathbf{Y}_q$. Clearly, with $\mathscr{H}$ being a very high-dimensional space, (4.3) cannot be solved directly. To overcome this limitation, the kernel trick is used where the original subspaces $\mathbf{Y}_q$ are never explicitly mapped to $\mathscr{H}$ (Mika et al., 1999). Instead, they are represented through a set of pairwise similarity comparisons based on a valid kernel function defined on the Grassmann manifold. The Grassmann kernel used in our approach is defined as (Hamm and Lee, 2008)

$$k(\mathbf{Y}_p, \mathbf{Y}_q) = \left\| \mathbf{Y}_p^T \mathbf{Y}_q \right\|_F^2, \tag{4.6}$$

with $\{\cdot\}_F$ denoting the matrix Frobenius norm and $\mathbf{Y}_p$ and $\mathbf{Y}_q$ being the orthonormal matrices representing the (spectral or temporal) subspaces of samples p and q. Using the Grassmann kernel in (4.6), (4.3) can be reformulated without explicitly computing $\mathbf{S}_b^\phi$ and $\mathbf{S}_w^\phi$ and the discriminant directions $\mathbf{w}$ can be analytically computed and used to project the spectro-temporal subspaces onto a lower-dimensional Euclidean space.[2] The final classification results presented in Section 4.3.4 are then obtained using LDA on these dimensionality-reduced subspaces.

It should be noted that as in KLDA, generally $C-1$ discriminant directions can be found, where C is the number of classes. Hence, by using $C-1$ discriminant directions, the dimensionality of the resulting Euclidean space can be reduced to $C-1$. For our two-class problem (e.g.,

[2] For details on reformulating (4.3) using the kernel trick and computing the discriminant direction $\mathbf{w}$, the reader is referred to Mika et al. (1999).

$C = 2$), the final resulting lower-dimensional Euclidean space is of dimension 1, therefore the final LDA classifier has only 1 parameter.

4.3 Experimental results

In this section, the performance of the proposed subspace-based approach for pathological speech detection is investigated and compared to state-of-the-art approaches.

As mentioned before, on the one hand, the temporal subspace-based detection method requires all speakers to utter the same speech material and uses the speech signals of a healthy speaker as a reference. On the other hand, spectral subspace-based and other state-of-the-art approaches do not have such a requirement. For a fair comparison of all approaches, in the following, we consider two scenarios. The first scenario is designed to compare the temporal subspace-based method and other methods by excluding a subset of the speech material and several healthy speakers before the evaluation. The second scenario is designed to compare the performance of all other approaches on the complete databases.

4.3.1 Evaluation protocols

The applicability and generalisability of the proposed approach are evaluated on two databases, i.e., the Spanish PC-GITA and French MoSpeeDi databases (cf. Section 2.2.1 and 2.2.2). The construction of the two previously explained scenarios on the two databases is done as follows.

Scenario 1

Due to computational limitations of DTW for time alignment when temporal subspaces are used, from each database we consider only 6 sentences for each speaker. After preprocessing (cf. Section 2.2), all sentences are concatenated and used to extract spectro-temporal subspaces and state-of-the-art features for each speaker (cf. Section 4.3.3).

Reference speakers for time-alignment As described in Section 4.2.1, computing temporal subspaces requires a reference speaker for time-alignment. To avoid introducing any bias, the considered reference speakers are not included in the training/testing sets of the databases. To analyze the sensitivity of the temporal subspace-based approach to the reference speaker selection, 5 and 10 randomly selected (healthy) reference speakers are considered for the PC-GITA and MoSpeeDi databases, respectively. Excluding the reference speakers from the databases and also maintaining the balance between the number of speakers in each class, we consider 45 PD patients (22 males, 23 females) and 45 healthy speakers (22 males, 23 females) from the PC-GITA database and 20 PD and ALS patients (14 males, 6 females) and 20 healthy speakers (10 males, 10 females) from the MoSpeeDi database to train and test the final classification using GDA. The validation strategy on the PC-GITA and MoSpeeDi databases for this scenario is a stratified speaker-independent 9-fold and 5-fold cross-validation, respectively.

The performance of the proposed temporal subspace-based approach using each reference speaker is computed, and the presented performance values in Section 4.3.4 for the temporal subspace-based approach represent the mean and standard deviation of this performance across different reference speakers. The same evaluation paradigm is also used for state-of-the-art approaches.

Scenario 2

To maintain comparability to the literature and across the different chapters in this thesis, we also evaluate the performance of the different approaches on the complete available databases. In such a scenario, the temporal subspace-based method is not applicable (due to requirements mentioned in the previous subsection). Hence, in this scenario we evaluate the spectral subspace-based method and the state-of-the-art approaches using all the speech material available for both databases, i.e., considering total 100 speakers in PC-GITA and 40 speakers in MoSpeeDi (cf. Section 2.2.1 and 2.2.2). The validation strategy on the PC-GITA and MoSpeeDi databases for this scenario is a stratified speaker-independent 10-fold and 5-fold cross-validation, respectively.

4.3.2 Algorithmic settings

Spectro-temporal subspaces are extracted on the logarithm of one-third octave band representations and MFCC representations. The one-third octave band representation is computed using $J = 15$ and a 32 ms Hamming window with 50% overlap. For the MFCC representation, similar framing parameters as in the octave band representation and 20 Mel filters are considered. Finally the first 15 MFCC coefficients are used (cf. Section 2.4).

As in Hamm and Lee (2008), a regularization parameter δ is used for GDA to avoid numerical issues and improve generalisability. Therefore, our subspace-based approach has two hyper-parameters, i.e., δ and the number of spectral basis vectors (d_s) or the number of temporal basis vectors (d_t). To select δ, d_s, and d_t, a grid-search with $\delta \in \{10^{-10}, \ldots, 10^{-1}\}$, $d_s \in \{1, \ldots, J\}$, and $d_t \in \{1, \ldots, J\}$ is performed using nested cross-validation in each training fold. The final δ, d_s, and d_t are selected as the ones yielding the highest mean accuracy on the training set.

4.3.3 State-of-the-art methods and baseline features

The proposed subspace-based approach is compared to using an SVM with a radial basis kernel function with state-of-the-art features such as MFCCs and the frequency-dependent shape parameter μ. Furthermore, we also evaluate using an SVM with one-third octave band representations, which has not been previously explored for this task. When using MFCCs, the feature vector is a 60–dimensional vector constructed by extracting 4 functionals, i.e., mean, standard deviation, kurtosis, and skewness of 15 MFCCs across time. When using the shape parameter μ, the feature vector is a 385–dimensional vector constructed as in Kodrasi and

Table 4.1 – Performance of the proposed (i.e., T-GDA and S-GDA) and state-of-the-art pathological speech detection methods (i.e., SVM using TF functionals and SVM using the sparsity parameter) on different databases using evaluation scenario 1.

		Spanish PC-GITA		French MoSpeeDi	
Method	Representation	Accuracy (%)	AUC	Accuracy (%)	AUC
		Proposed			
T-GDA	MFCC	79.1 ± 4.2	0.87 ± 0.03	$\mathbf{90.5} \pm 3.3$	$\mathbf{0.94} \pm 0.02$
T-GDA	Octave	$\mathbf{82.9} \pm 1.9$	$\mathbf{0.90} \pm 0.01$	81.5 ± 4.6	0.89 ± 0.03
S-GDA	MFCC	78.9	0.84	75.0	0.78
S-GDA	Octave	61.1	0.71	65.0	0.66
		Baseline			
SVM on functionals	MFCC	72.2	0.81	65.0	0.77
SVM on functionals	Octave	69.0	0.75	67.5	0.75
SVM	μ	67.8	0.77	72.5	0.77

Bourlard (2020). For both considered feature vectors, to select the soft margin constant C and the kernel width γ of the SVM, a grid search with $C \in \{10^{-2}, \ldots, 10^{4}\}$ and $\gamma \in \{10^{-4}, \ldots, 10^{2}\}$ is performed using nested cross-validation in each training fold. The final C and γ are selected as the ones yielding the highest mean accuracy on the training set.

It should be noted that we also investigated the performance using an SVM with other state-of-the-art acoustic features such as f_0, jitter, shimmer, and HNR. However using such features did not perform well for this task (Janbakhshi et al., 2020; Kodrasi and Bourlard, 2020). Therefore we decided to report the result of the acoustic features which have shown more promising results in the literature (e.g., MFCCs and sparsity parameters).

4.3.4 Results

Table 4.1 presents the accuracy and AUC of the considered pathological speech detection approaches on the considered databases using evaluation scenario 1, with bold entries indicating the maximum performance. The proposed spectral and temporal subspace-based approaches are denoted by S-GDA and T-GDA, respectively. The SVM methods refer to state-of-the-art (baseline) approaches. For the proposed temporal subspace-based approach on the PC-GITA and MoSpeeDi databases, besides the mean performance, the standard deviation of the performance across different reference speakers are also presented (cf. Section 4.3.1).

Several observations can be made based on the presented results. First, it can be observed that the proposed subspace-based approach using temporal subspaces yields better performance than using spectral subspaces and outperform all other state-of-the-art approaches for both considered databases independently of the used speech representation. Hence, it can be said

Table 4.2 – Performance of the proposed (i.e., S-GDA) and state-of-the-art pathological speech detection methods (i.e., SVM using TF functionals and SVM using the sparsity parameter) on different databases using evaluation scenario 2.

		Spanish PC-GITA		French MoSpeeDi	
Method	Representation	Accuracy (%)	AUC	Accuracy (%)	AUC
Proposed					
S-GDA	MFCC	**77.0**	**0.82**	67.0	0.72
S-GDA	Octave	65.0	0.69	70.0	0.65
Baseline					
SVM on functionals	MFCC	65.0	0.68	52.5	0.52
SVM on functionals	Octave	70.0	0.76	**72.5**	**0.8**
SVM	μ	71.0	0.78	67.5	0.70

that the characterization of temporal patterns has higher discriminative power for subspace-based healthy and pathological speech discrimination than the characterization of spectral patterns. Further, observing the low standard deviation of the performance of the temporal subspace-based approach suggests that this approach is not highly sensitive to the reference speaker selection. Although an optimal reference speaker can be chosen based on nested cross-validation on the training data, we did not attempt to find the best reference speaker for alignment. This is a fair comparison for scenarios where many reference speakers are not available. Second, it can be observed that the spectral subspace-based approach using MFCC representations performed better than other state-of-the-art approaches. Comparing the two speech representations on both databases, we observe that baseline approaches, i.e., SVM on functionals of MFCC and octave-band representations yield comparable performance while using spectral subspace-based approach (i.e., S-GDA) only improves the performance using MFCC representations. It should be noted that in both approaches, the temporal information is ignored and only spectral information is considered.

In summary, the proposed temporal subspace-based method outperforms the state-of-the-art methods achieving better performance on both considered databases and both considered feature representations. However, such an approach is applicable only in phonetically-balanced scenarios.

Table 4.2 presents the performance of the considered pathological speech detection approaches for scenario 2, where we compare different spectral approaches to establish the baseline performance on all available data from the two databases. First, it can be observed that S-GDA using MFCC representations yields the best performance on the PC-GITA database while using an SVM with octave band representations yields the best performance on the MoSpeeDi database. Nevertheless, the performance of the S-GDA using MFCC representations on the MoSpeeDi database is not significantly lower than the performance of best performing

approach on this database, i.e., using an SVM with octave band representations.

In summary, when considering all speakers and available speech material, the proposed S-GDA with MFCC representations yields better or comparable performance than state-of-the-art approaches.

4.4 A note on extending linear subspace-based analysis

In this chapter, temporal subspaces when compared to spectral subspaces, are shown to be more successful in characterizing pathological speech. However, constructing the temporal subspaces requires time-alignment limiting its application to only phonetically-balanced scenarios. In further analysis, we have attempted to further improve the performance of the spectral subspace-based discriminant analysis approach, since it does not require time-alignment and is applicable to phonetically-unbalanced scenarios. While in the current chapter only linear subspace-based discriminant analysis was used, in this further analysis we characterized nonlinear spectral patterns of speech using nonlinear subspaces obtained by nonlinear PCA benefiting from the kernel trick (kernel PCA) (Schölkopf et al., 1997). Automatic pathological speech detection was then achieved by nonlinear (kernel) subspace-based discriminant analysis. To partly incorporate short-time temporal information in the constructed subspaces, we considered Hankel representations (Zhao and Liu, 2004; Ku et al., 1995; Luo et al., 2019). Hankel representations are modified TF representations, where each column vector was constructed by concatenating several temporally consecutive spectral vectors from the original TF representation, resulting in a higher-dimensional vector. By representing healthy and pathological speakers by nonlinear (kernel) subspaces, the kernelized GDA (KGDA) (Wang and Shi, 2009) was used for the final classification of speakers. We achieved promising results showing that using nonlinear subspaces modeling Hankel representations was superior to using linear spectral subspaces modeling TF representations for the speech pathology detection task. However, finding an efficient way to select the optimal number of kernel basis vectors to construct the nonlinear subspaces remained an unsolved issue. We omitted the results of our nonlinear subspace analysis in this chapter, however, we would like to note that this topic should be further investigated.

4.5 Summary

To automatically discriminate between pathological and healthy speech, in this chapter we have proposed a subspace-based approach representing speakers through spectral or temporal subspaces spanned by the dominant spectral or temporal basis vectors of the feature representation of speech. Prior to constructing the temporal subspaces, it has been proposed to time-align signals to a reference representation using DTW. The spectral and temporal basis vectors are extracted using the SVD. Since speakers are represented through subspaces, it has been proposed to apply subspace-based discriminant analysis to automatically discriminate

between pathological and healthy speakers. Experimental results on two databases have shown that compared to spectral subspaces, temporal subspaces are more successful in characterizing pathological speech. In addition, it has been shown that the proposed subspace-based approach using temporal subspaces outperforms using an SVM with state-of-the-art features for pathological speech detection. A limitation of the temporal subspace-based approach is that it relies on having access to utterances with the same phonetic content from both healthy and pathological speakers. However, when the phonetic content of the available utterances from speakers differs, the proposed subspace-based approach using spectral subspaces can be used instead.

5 Deep learning for automatic pathological speech detection

In this chapter we propose different deep learning-based approaches as alternatives to classical machine learning-based approaches to automatically discriminate between pathological and healthy speech. First, we present our approach based on a pairwise distance-based CNN. Then, we describe our proposed supervised speech representation frameworks using CNNs to achieve pathological speech detection. The applicability of the proposed approaches in this chapter is experimentally investigated across databases and compared to many baseline frameworks.

5.1 Introduction

In Chapter 4 we evaluated the performance of classical machine learning-based methods for pathological speech detection and demonstrated that our proposed temporal subspace-based learning method outperforms the state-of-the-art approaches. As mentioned in Chapter 2, handcrafting appropriate acoustic features is a crucial step in such machine learning-based approaches, however, such features may fail to characterize abstract (but similarly important) acoustic cues that can further assist in differentiating pathological speech from healthy speech. Therefore, deep learning approaches can be used for learning high-level abstract speech representations for such a task. While deep learning approaches have dramatically improved the state-of-the-art in many speech processing applications, their advantages are yet to be established in the field of pathological speech assessment (Cummins et al., 2018).

In this chapter we focus on extracting high-level representations of speech using data-driven deep learning approaches. The main challenge in successfully learning such representations is being able to alleviate overfitting issues and also to guide the networks to learn robust and relevant features for pathological speech detection while using the small amount of pathological training data that is typically available. Based on these motivations, in this chapter, we propose two different deep learning frameworks using CNNs for automatic pathological speech detection.

5.1.1 Pairwise distance-based convolutional neural networks

For speech pathology detection, deep learning-based approaches are usually based on modeling short segments of speech since splitting speech signals into short segments (e.g., 160 ms) increases the number of training samples per speaker. However, such short segments do not always exhibit pathological characteristics and the CNNs are not guided to ignore speaker variabilities that are unrelated to a speech disorder. Considering that larger segments of speech (e.g., word-level) can be expected to better reflect pathological characteristics, we need a framework which can cope with the limited number of word utterances in training data. Therefore, we propose pairwise training CNNs using distance matrices constructed from representations of words. Pairwise training is advantageous for limited training data, since it guides the network to extract features that are discriminative of pathological speech while being robust to other unrelated speaker variabilities. In the literature, pairwise training has been exploited before for such a task (Bhati et al., 2019), however, the used network in Bhati et al. (2019) is phonetic content specific, i.e., it requires training different networks for different utterances. While our proposed system benefits from pairwise training, a single network can be used for different utterances, since it operates on distance matrices instead of operating directly on pairs of input data as in Bhati et al. (2019).

Inspired by the CNN-based query detection system in Ram et al. (2020), we consider utterances from healthy speakers as reference representations and we propose to compute frame-level distance matrices between these reference representations and phonetically-balanced test representations. We hypothesize that when the test speaker is healthy, the pattern of the distance matrix between the test and reference (i.e., healthy) representations is different (i.e., it is expected to be more quasi-diagonal) than when the test speaker is pathological. This distance pattern can be used as the input to a CNN-based binary classifier, which then categorizes it as an example from a healthy speaker (i.e., the distance pattern arises from comparing a healthy utterance to the reference representation) or as an example from a pathological speaker (i.e., the distance pattern arises from comparing a pathological utterance to the reference representation). Such a system can operate on any user-defined representation of utterances such as the STFT representation. Instead of the STFT representation as in Vasquez et al. (2017), we propose to use articulatory posteriors (APs). The use of APs is motivated by their potential to characterize articulation deficits in speech pathologies, their robustness to noise, and their multilingual and cross-lingual portability (Rasipuram and Magimai.-Doss, 2016).

Figure 5.1 shows two examples of distance matrices computed from AP representations of a sample utterance belonging to a test pathological (dysarthric) speaker and a test healthy speaker. The reference representation used in both distance matrices is the same and is from a healthy speaker (different from the test healthy speaker). As can be observed, the distance matrix corresponding to the test healthy speaker (shown in Figure 5.1b) has different (e.g., more quasi-diagonal) patterns compared to the distance matrix corresponding to the test pathological speaker (shown in Figure 5.1a).

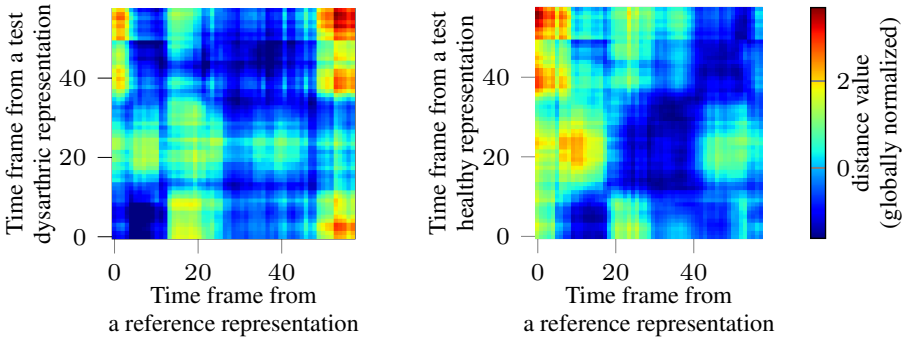


Figure 5.1 – Distance matrices computed from AP representations of a sample utterance from (a) a test pathological and a reference speaker (b) a test healthy and the reference speaker. The distance matrix corresponding to the test healthy speaker has different (e.g., more quasi-diagonal) patterns compared to the distance matrix corresponding to the test pathological speaker.

Although a CNN can directly operate on distance matrices computed from user-defined representations of utterances (e.g., STFT or AP), these user-defined representations might not be optimal for healthy and pathological speech detection. To ensure that distance matrices are computed on optimal representations for our task, we propose to incorporate a front-end feature extraction layer into the network prior to computing the distance matrices. The front-end feature extraction, the distance matrix computation, and the final healthy and pathological speech detection layers are jointly optimized in an end-to-end learning framework.

5.1.2 Supervised speech representation learning

As previously mentioned, our first proposed approach exploits pairwise training while using a single network for different utterances, which guides the network to extract features that are discriminative of speech pathologies while being robust to other unrelated speaker variabilities. However, such an architecture relies on having access to utterances with the same phonetic content from both healthy and pathological speakers.

Recently it has been proposed to learn high-level (but not necessarily robust and discriminative as explained in the following) representations through unsupervised auto-encoders operating on phonetically unmatched speech segments (Vasquez-Correa et al., 2020; Karan et al., 2020b). The extracted representations are then used as input for training pathological speech classifiers. Unsupervised representation learning based on auto-encoders yields representations that are designed to reconstruct the input. Consequently, there is no guarantee that these learned representations are robust to pathology-unrelated cues such as acoustic information about the speaker's identity. In addition, there is no guarantee that these representations are discriminative for pathology detection.

To tackle these issues, in this thesis we propose methods to extract robust and discriminative representations from speech exploiting supervised auto-encoders.

Using a single encoder, first we propose to supervise the representation learning process such that only speaker-invariant information is retained in the obtained learned feature representation. This is achieved through training an adversarial network by jointly minimizing the auto-encoder reconstruction loss and the performance of a (healthy) speaker identification (ID) task. The prominence of speaker variabilities unrelated to speech pathology in such representations will be limited, and hence, it can be expected that the performance of pathology detection can be improved. Suppressing unrelated speaker variabilities from representations in an adversarial training framework has been recently shown to improve the performance for different classification tasks such as speech emotion classification, phoneme/senone discrimination, and speaker de-identification (Meng et al., 2018; Li et al., 2020; Higuchi et al., 2019; Espinoza-Cuadros et al., 2020). Second, to ensure that the learned representations retain pathological speech discriminative information, using the same architecture (i.e., with a single encoder) we propose to train the representation layer by jointly minimizing the auto-encoder reconstruction loss and maximizing the performance of pathological speech detection. in Le et al. (2018) it has been shown that such supervised auto-encoders typically do not harm the performance compared to a standard neural network, since the incorporation of the reconstruction loss into the training procedure acts as a regularisation method. It should be noted that such a joint training procedure to learn discriminative representations for pathological speech detection has been investigated in Korzekwa et al. (2019). However, in Korzekwa et al. (2019) two encoders are used, i.e., an audio and a text encoder. Differently from Korzekwa et al. (2019) and inline with unsupervised representation learning in Vasquez-Correa et al. (2020); Karan et al. (2020b), a single encoder is used in our framework.

Our experimental results on the larger considered database (cf. Section 5.3.3) show that using the supervised speaker-invariant representations can be as effective as using the supervised pathology discriminative representations for improving the performance when using unsupervised learned representations. However, due to the difficulty of training the adversarial networks for obtaining the speaker-invariant representation mentioned before, we also propose a dual representation learning framework based on a feature separation such that the speaker identity-related features are isolated without using any adversarial training. Among the limitations of adversarial training is that they might suffer from oscillating and unstable training where convergence can be evaded (Sha and Lukasiewicz, 2021). Furthermore, in Moyer et al. (2018), specific failure modes of adversarial training are also argued. Our adversarial-free proposed system consists of two encoders: the first encoder generates a bottleneck feature representation containing speaker identity information supervised by maximizing the performance of a speaker ID auxiliary task; whereas the second encoder generates speaker-invariant feature representations. To reduce the impact of speaker identity cues in the second encoded bottleneck representation, the mutual information (MI) between the two encoded bottleneck representations is minimized. To avoid loss of information embedded within the representative features, a decoder fed by both encoded features is simultaneously

trained to minimize the reconstruction loss. Such a training procedure reducing the dependency between the two learned representations yields a feature representation (generated by the second encoder) that contains less cues about speaker identities, and therefore, is a more robust representation for pathological speech detection.

Estimating the MI between high dimensional continuous variables is a difficult task, therefore, different estimators of the upper bound and lower bound of MI using neural network architectures have been proposed and considered for optimization (Belghazi et al., 2018; Cheng et al., 2020). Optimizing MI estimators to separate the latent representations into (ideally) independent components while avoiding adversarial training has been exploited for different applications, such as unsupervised domain adaptation for image classification tasks, voice style transfer (voice conversion), and speech synthesis (Cheng et al., 2020; yao Hu et al., 2020; Yuan et al., 2021). In previous work, adversarial methods are still used for feature separation while the MI minimizer is incorporated in the training to strengthen the feature separation performance, e.g., for cross-lingual TTS synthesis and learning domain-agnostic representations for computer vision tasks (Peng et al., 2019; Xin et al., 2021). In addition, in a domain adaptation framework to improve cross-domain pathological speech detection in Wang et al. (2021), MI minimization along with domain adversarial training is used to separate speech pathology discriminative information from domain-related information. Contrary to the state-of-the-art literature, in this thesis we propose to obtain a speaker-invariant representation for pathological speech detection using a feature separation framework relying on MI minimization criteria without using any adversarial training. To the best of our knowledge, the applicability of such a framework to improve pathological speech detection has not been explored before.

This chapter is organized as follows. Section 5.2.1 presents the approach based on pairwise distance-based CNNs. The performance of this approach is evaluated in Section 5.2.2. Supervised speech representation learning approaches using a single encoder to obtain speaker-invariant and/or pathological speech discriminative representations are presented in Section 5.3.1. The alternative approach to obtain speaker-invariant representation using a feature separation framework instead of adversarial training is presented in Section 5.3.2. Experimental results and discussions for all supervised representation learning methods are presented in Section 5.3.3. Finally, we summarize our findings in Section 5.4.

5.2 Pairwise distance-based convolutional neural networks

5.2.1 Proposed approach

Fig. 5.2 depicts a schematic representation of the proposed pairwise distance-based pathological speech detection CNN. As shown in this figure, the input to the system consists of pairs of reference and test representations of utterances. We follow the same procedure as in Dubagunta and Magimai-Doss (2019) to extract AP features for the representations of utter-

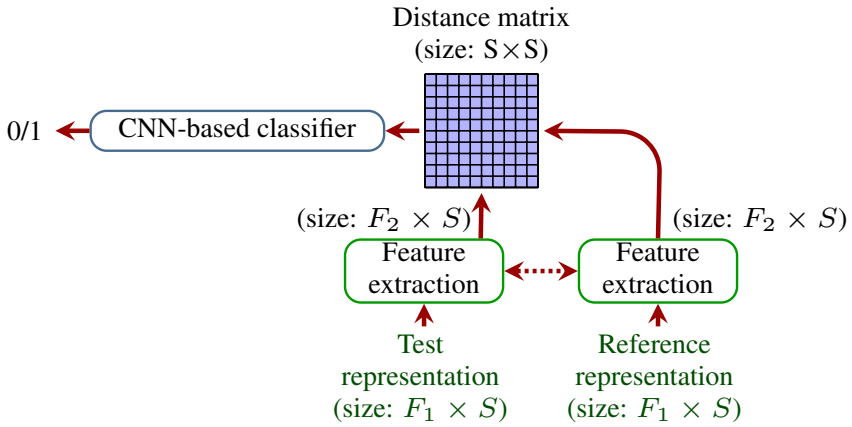


Figure 5.2 – Block diagram of the proposed pairwise distance-based pathological speech detection CNN. For this approach, pairs of phonetically-matched reference and test utterance representations are considered. Given such pair inputs to the system, features are first extracted, then distance matrices are computed, which are then further evaluated by a CNN-based classifier to predict whether the test representation is from a pathological or healthy speaker. The two feature extraction blocks share the same set of parameters.

ances (cf. Section 5.3.3). These representations are transformed through a feature extraction block prior to computing the distance matrix. The distance matrix is then considered as an image by a CNN-based classifier as in a standard binary image classification task. The complete architecture is optimized in an end-to-end framework to achieve pathological speech detection.

In the following, we present details on the different components of the proposed system, i.e., i) the front-end feature extraction, ii) the distance matrix computation, and iii) the CNN-based classifier.

Front-end feature extraction

We consider pairs of phonetically-balanced AP representations of utterances from two speakers; one utterance being a reference representation from a healthy speaker and the other utterance being from a test (healthy or pathological) speaker. Let us denote by $\mathbf{R}$ the $(F_1 \times M)$–dimensional reference representation, with F_1 being the number of AP features and M being the number of time frames in the reference representation. Similarly, let us denote by $\mathbf{T}$ the $(F_1 \times N)$–dimensional test representation, with N being the number of time frames in the test representation. To be able to handle variable-length inputs, we fix the length of all representations to a predetermined (user-defined) size S as in Ram et al. (2020). Representations with more time frames than S, i.e., $M > S$ or $N > S$, are down-sampled by deleting time frames in regular intervals. Representations with less time frames than S, i.e., $M < S$ or $N < S$, are padded at the beginning and the end with time frames filled with a constant value. The constant value is arbitrarily set to the maximum value in the representation. We denote the

Table 5.1 – Front-end feature extraction architecture.

Layer	Description
Input	Size: (1xF_1xS): input speech representation
Conv1d + Relu	Channel: in=1, out=32, Filter: F_1x1, Stride: 1

resized reference and test representations by $\mathbf{R}_s$ and $\mathbf{T}_s$ and hypothesize that they contain similar (healthy or speech pathology-related) cues as in the original representations $\mathbf{R}$ and $\mathbf{T}$.

The front-end feature extraction block transforms the $(F_1 \times S)$–dimensional representations $\mathbf{R}_s$ and $\mathbf{T}_s$ into $(F_2 \times S)$–dimensional representations. To this end, we use a 1D convolution layer with F_2 channels such that the F_1–dimensional AP feature vectors for each time frame are transformed into F_2–dimensional feature vectors. Since this layer is jointly optimized with the distance matrix computation and the CNN-based classifier (cf. following subsections) in an end-to-end framework, it can be expected that the transformed $(F_2 \times S)$–dimensional representations are more discriminative representations for the pathological speech detection task.

The architecture of the front-end layer is summarized in Table 5.1, where we have used $F_2 = 32$. It should be noted that the parameters of the front-end feature extraction layer to compute both test and reference feature representations are the same (cf. Fig. 5.2).

Distance matrix computation

The distance matrix is computed from the representations at the output of the feature extraction block. Let us denote the reference representation after feature extraction by $\hat{\mathbf{R}} = [\mathbf{r}_1, \ldots, \mathbf{r}_S]$, with $\mathbf{r}_i$, $i = 1, \ldots, S$, being the F_2–dimensional feature vector at time frame i. Similarly, the test representation after feature extraction is denoted by $\hat{\mathbf{T}} = [\mathbf{t}_1, \ldots, \mathbf{t}_S]$, with $\mathbf{t}_j$, $j = 1, \ldots, S$, being the F_2-dimensional feature vector at time frame j. The frame-level distance matrix $\mathbf{D}$ between the representations $\hat{\mathbf{T}}$ and $\hat{\mathbf{R}}$ is an $(S \times S)$–dimensional matrix, where the (i, j)–th entry is computed as the distance d between $\mathbf{t}_i$ and $\mathbf{r}_j$, i.e.,

$$\mathbf{D}_{i,j} = d(\mathbf{t}_i, \mathbf{r}_j). \tag{5.1}$$

To compute $\mathbf{D}$ within the proposed end-to-end framework, Euclidean distance is used, i.e., $d(\mathbf{t}_i, \mathbf{r}_j) = ||\mathbf{t}_i - \mathbf{r}_j||$. Since the reference representation $\hat{\mathbf{R}}$ always belongs to a healthy speaker, we expect the pattern of the so-computed distance matrix $\mathbf{D}$ to be more quasi-diagonal (i.e., contain more zeros on the diagonal due to similar $\mathbf{t}_i$ and $\mathbf{r}_j$) when the test representation $\hat{\mathbf{T}}$ belongs to a healthy speaker than when it belongs to a pathological speaker.

Table 5.2 – Architecture of the proposed CNN-based classifier operating on pairwise distance matrices.

Layer	Description
Input	Size: (1xSxS) input distance matrix
Conv2d + Relu	Channel: in=1, out=16, Filter: 10x10, Stride: 1
Maxpool2d	Channel: in=16, out=16, Filter: 2x2, Stride: 2
Conv2d + Relu	Channel: in=16, out=16, Filter: 10x10, Stride: 1
Maxpool2d	Channel: in=16, out=16, Filter: 2x2, Stride: 2
Dropout	Probability: 0.5
FC + Relu	Input: 784, Output: 128
FC + Softmax	Input: 128, Output: 2

CNN-based classifier with pairwise distance matrices

The distance matrices computed in the previous subsection serve as input to our CNN classifier. As summarized in Table 5.2, the CNN classifier consists of two 2D convolutional layers, followed by two Maxpooling and two fully connected (FC) layers. To prevent overfitting, dropout is employed during training. The label for each distance matrix fed into the CNN classifier is the label of the test speaker (healthy or pathological) used for the distance matrix computation.

The classifier is trained using distance matrices computed from all phonetically-matched pairs of test and reference representations in the training set. As mentioned in Section 5.1.1, a single network can be used for different utterances since the CNN operates on distance matrices instead of pairs of input data as in Bhati et al. (2019). To evaluate an utterance from an unseen test speaker, we pair it to its phonetically-matched counterpart from many reference speakers in the training set and compute multiple distance matrices. All available distance matrices are then independently processed by the CNN classifier, and the final decision for the unseen test speaker is made by applying soft voting on all CNN prediction scores for all available distance matrices from that speaker.

5.2.2 Experimental results

Baseline networks

To demonstrate the advantages of the proposed approach, the following two baseline systems B-CNN$_1$ and B-CNN$_2$ are considered.

B-CNN$_1$. We have implemented a baseline CNN adapted from Vasquez et al. (2017), which is trained on log magnitude of STFT representations of short (i.e., 160 ms) segments of speech with 50% overlap. Each segment is labeled as healthy or pathological depending on the label of the complete signal. The STFT representations are computed using 10 ms Hanning windows

Table 5.3 – Architecture of the baseline B-CNN$_1$ adapted from Vasquez et al. (2017).

Layer	Description
Input	Size: (1xFx16); F: dimension of input representation
Conv1d + Relu	Channel: in=1, out=32, Filter: Fx1, Stride: 1
Conv1d + Relu	Channel: in=32, out=16, Filter: 1x4, Stride: 1
Dropout	Probability: 0.5
FC + Relu	Input: 208, Output: 128
FC + Softmax	Input: 128, Output: 2

without overlap, resulting in 129 frequency bins for each time frame. The final decision for an unseen speaker is made by applying soft voting on the segment-level CNN prediction scores. To demonstrate the advantage of using AP representations instead of STFT, such a baseline CNN is also trained on the logarithm of AP representations. The architecture of this baseline system is summarized in Table 5.3.

B-CNN$_2$. To further establish the advantages of the proposed end-to-end CNN framework (which uses a front-end feature extraction layer), a second baseline is implemented where the proposed CNN-based classifier in Section 5.2.1 is trained on distance matrices computed directly from AP representations (i.e., without using the front-end feature extraction layer). To compute such distance matrices, Kullback-Leibler divergence is used as the local distance measure in (5.1). The architecture of this baseline system is the same as in Table 5.2.

Training and evaluation

The applicability and generalisability of the proposed approach is evaluated on the considered database, i.e., the Spanish PC-GITA database and the French MoSpeeDi database (cf. Section 2.2.1 and 2.2.2). As mentioned before, our pairwise distance-based network operates on pairs of word utterances, therefore here we use word utterances from each database. Similarly to the previous chapter, the validation strategy on the PC-GITA and MoSpeeDi databases is a stratified speaker-independent 10-fold and 5-fold cross-validation, respectively. As described in Section 2.4.5, AP features are extracted as in Dubagunta and Magimai-Doss (2019), and by concatenating all extracted APs, $F_1 = 53$ features per time frame are obtained.

In each training fold, a development fold with the same size as the test fold is set aside for early-stopping. Z-score normalization is applied to all input representations. All networks are trained using the stochastic gradient descent (SGD) algorithm and the cross-entropy loss. The batch size is 256, and the initial learning rate is 0.05. The learning rate is divided by 5 each time the loss on the development set does not decrease for 5 consecutive iterations. The training is stopped either after 100 epochs or after the learning rate has reached the value 10^{-6}.

Random weight initialization is used for the baselines B-CNN$_1$ and B-CNN$_2$. The weights on the first convolution layer of the trained baseline B-CNN$_1$ are used to initialize the front-end

Table 5.4 – Mean and standard deviation of the speech pathology detection accuracy [%] and AUC score using the baseline B-CNN$_1$ with STFT and AP representations on the PC-GITA and MoSpeeDi databases.

	Spanish PC-GITA		French MoSpeeDi	
Input representation	Accuracy (%)	AUC	Accuracy (%)	AUC
STFT	53.7 ± 3.3	0.56 ± 0.03	52.5 ± 0.0	0.64 ± 0.02
AP	72.0 ± 0.8	0.75 ± 0.00	60.8 ± 3.1	0.73 ± 0.03

feature extraction layer of the proposed end-to-end CNN. The weights of the trained baseline B-CNN$_2$ are used to initialize the classifier layers of the proposed end-to-end CNN.

The number of total samples (training/testing) available for the different considered networks is as follows. Using the STFT representation for B-CNN$_1$ results in 17383 (PC-GITA) and 25197 (MoSpeeDi) segments. Using the AP representation for B-CNN$_1$ results in 17368 (PC-GITA) and 25907 (MoSpeeDi) segments. The number of distance matrices computed from all pairs of reference and test AP representations for B-CNN$_2$ and the proposed CNN is 96000 (PC-GITA) and 25920 (MoSpeeDi).

To reduce the impact of the random seed on the final model parameters, we have trained all networks with multiple different random seeds. The reported performance measures are the mean and standard deviation of the performance obtained by models trained using different seeds.

Results

Table 5.4 presents the classification accuracy and AUC values obtained using B-CNN$_1$ on STFT and AP representations for both considered databases. It can be observed that the AP representation yields a better performance than the STFT on both databases, with a particularly significant improvement observed for the PC-GITA database. These results are to be expected given the advantages of articulatory modeling of speech using AP as described in Section 5.1.1.

It should be noted that the CNN proposed in Vasquez et al. (2017) was trained on the PC-GITA database using speech segments centered at transitions between voiced and unvoiced regions. However, although not presented here, using such segments did not result in a better performance than the performance presented in Table 5.4. Further, it should be noted that Vasquez et al. (2017) uses more recordings than the word recordings we have used here. To ensure that the conclusions derived in this chapter on the advantages of the proposed approach as opposed to B-CNN$_1$ are still valid even when more recordings are available, we have investigated the performance of B-CNN$_1$ using AP representations on both databases when all available recordings are used (rather than just words).

Table 5.5 – Mean and standard deviation of the speech pathology detection accuracy [%] and AUC score using the baseline B-CNN$_1$ and B-CNN$_2$ and the proposed pairwise distance-based approach with a front-end feature extraction layer on the PC-GITA and MoSpeeDi databases.

	Spanish PC-GITA		French MoSpeeDi	
CNN	Accuracy (%)	AUC	Accuracy (%)	AUC
Baseline B-CNN$_1$	72.0 ± 0.8	0.75 ± 0.00	60.8 ± 3.1	0.73 ± 0.03
Baseline B-CNN$_2$	68.3 ± 0.7	0.78 ± 0.01	70.8 ± 2.3	0.77 ± 0.00
Proposed	$\mathbf{77.7} \pm 0.5$	$\mathbf{0.83} \pm 0.01$	$\mathbf{76.7} \pm 4.2$	$\mathbf{0.84} \pm 0.02$

Using all available recordings and the AP representation for B-CNN$_1$ results in 74762 (PC-GITA) and 54626 total available segments. In this case, B-CNN$_1$ yields accuracy and AUC values of 73.33% and 0.78 on the PC-GITA database and 60.0% and 0.75 on the MoSpeeDi database. When comparing these results to the ones obtained using only word recordings (cf. entries for AP representations in Table 5.4), we observe that increasing the used speech material does not significantly improve the pathological speech detection performance of B-CNN$_1$. In summary, the presented results demonstrate the advantage of using AP representations as opposed to the STFT representations used in Vasquez et al. (2017). In the following, the performance of both baseline systems B-CNN$_1$ and B-CNN$_2$ and of the proposed end-to-end CNN is compared when AP representations are used.

Table 5.5 presents the classification accuracy and AUC values of the baseline systems B-CNN$_1$ and B-CNN$_2$ and of the proposed approach on both databases. Bold entries indicate the maximum performance for each database. It can be observed that the proposed pairwise distance-based CNN with front-end feature extraction outperforms both considered baselines in terms of both performance measures on both databases. Comparing the difference in performance between the proposed framework and B-CNN$_2$ shows that incorporating a feature extraction front-end significantly improves the performance in comparison to computing distance matrices directly on AP representations. Analyzing the learned representations from the feature extraction front-end remains a topic for future investigation.

In summary, the presented results show that the proposed pairwise distance-based CNN with a front-end feature extraction layer is successfully applicable to the pathological speech detection task. Although a small number of utterances per speaker are used, the proposed approach outperforms CNN-based baseline systems for both databases.

Before comparing the results obtained by our pairwise distance-based CNN to the baseline classical machine learning-based approaches in Chapter 4, it should be noted that our pairwise distance-based CNN operates only on word utterances, while for the baseline classical machine learning-based approaches in Chapter 4 all speech material from speakers is used (cf. entries regarding the baseline in Table 4.2). Bearing in mind such differences, it can be observed that pairwise distance-based CNN performs better than classical baseline

approaches.

Both pairwise distance-based CNN and our previously proposed temporal subspace-based approach (cf. Section 4) rely on using utterances with the same phonetic content from both healthy and pathological speakers. However, as mentioned before, pairwise distance-based CNN operates only on word utterances while the temporal subspace-based approach operates on sentence-level speech signals. By considering the performance of the temporal subspace-based approach (cf. entries regarding T-GDA in Table 4.1) while bearing in mind the differences in the training speech material for both approaches, we observe that distance-based CNN has not performed better. As mentioned before, temporal subspace approach uses the long-term discriminative cues embedded in the speech signals while most CNN systems use short-term acoustic cues. Such long-term discriminative cues might be more powerful indicators of pathological speech than short-term cues. Therefore, generally, this can be a bottleneck in the performance of the CNN-based systems operating on short-term cues.

5.3 Supervised speech representation learning

In this section, we present two proposed supervised representation learning frameworks for the pathological speech detection task. The first framework aims at learning a single representation, while in the second framework, dual representation learning motivated by feature separation is proposed.

5.3.1 Proposed supervised single representation learning

Figure 5.3 illustrates the proposed representation learning for pathological speech detection using an auto-encoder (single encoder and decoder) and two auxiliary modules, i.e., an adversarial speaker ID module and a pathological speech classifier module. To obtain a speaker identity-invariant representation, the auto-encoder can be jointly trained with the speaker ID task in an adversarial manner. To obtain a pathological speech discriminative representation, the auto-encoder can be jointly trained with the pathological speech classifier. To obtain a speaker identity-invariant and pathological speech discriminative representation, the auto-encoder can be jointly trained with both auxiliary tasks.

Auto-encoder

Following a similar framework adapted from Vasquez-Correa et al. (2020), we consider a CNN-based auto-encoder to compute low-dimensional representations from chunks of speech TF representations. TF representations are encoded with three convolutional layers (filter size: 6×6, stride: 1), with the number of feature maps on each layer being twice the number of feature maps on the previous layer (starting with 32 maps in the first layer). Each convolutional layer is followed by max-pooling (filter size: 3×3, stride: 3), batch normalization, and ReLU

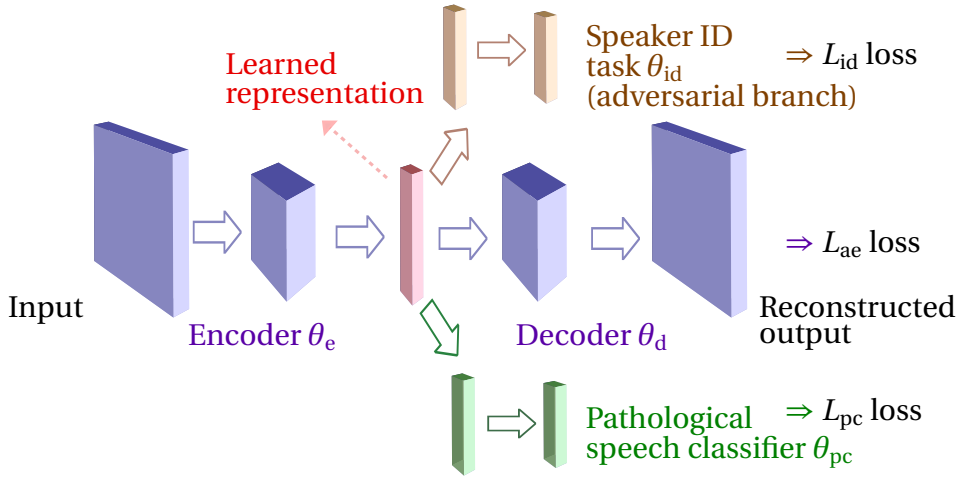


Figure 5.3 – Proposed supervised single representation learning for pathological speech detection using an auto-encoder and auxiliary tasks. The auto-encoder is jointly trained with the auxiliary speaker ID task and/or with the auxiliary pathological speech classifier.

activation functions. The output of the last convolutional layer is further processed with a fully connected layer (with 128 hidden units) to form the final feature representation, i.e., bottleneck representation, of size 128. The bottleneck representation is decoded into a reconstructed version of the input TF representations by the decoder. The decoder components are stacked in reverse order of the encoder components, where transposed convolutional and interpolation layers are used instead of convolutional and max-pooling layers. In the remainder of this section, the parameters of the encoder and decoder are denoted by θ_e and θ_d respectively.

Speaker identity-invariant representation with adversarial training

To learn representations robust to speaker variabilities unrelated to pathological speech, i.e., speaker identity, the bottleneck representation of the auto-encoder in the previous subsection is connected to a speaker ID module. The architecture of this module is adapted from the final classifier used in Vasquez-Correa et al. (2020) and consists of two fully connected layers with 64 hidden units each, a ReLU activation function after the first layer, and a Softmax activation function after the final (i.e., second) layer. The number of output units, i.e., the number of units in the final layer, is the same as the number of speakers used for the speaker ID task (cf.

Section 5.3.3). To avoid over-fitting, a dropout layer with a rate of 0.2 is included between the bottleneck layer and the speaker ID module. The parameters of this module are denoted by θ_{id}.

To obtain a compact representation where the information related to the speaker identity is minimized, we use adversarial training by minimizing the auto-encoder reconstruction loss L_{ae} such that a low reconstruction error is achieved while maximizing the speaker ID loss L_{id} such that a low speaker ID accuracy is achieved. Adversarial training is achieved through the min-max optimization objective

$$(\hat{\theta}_{\text{e}}, \hat{\theta}_{\text{d}}, \hat{\theta}_{\text{id}}) = \arg\min_{\theta_{\text{e}}, \theta_{\text{d}}} \arg\max_{\theta_{\text{id}}} E(\theta_{\text{e}}, \theta_{\text{d}}, \theta_{\text{id}}), \tag{5.2}$$

with

$$E(\theta_{\text{e}}, \theta_{\text{d}}, \theta_{\text{id}}) = L_{\text{ae}}(\theta_{\text{e}}, \theta_{\text{d}}) - \lambda L_{\text{id}}(\theta_{\text{e}}, \theta_{\text{id}}), \tag{5.3}$$

where λ is the trade-off parameter between the auto-encoder and the adversarial loss functions (cf. Section 5.3.3). In practice, the optimal parameters in (5.3) are approximated using an alternating training procedure, where in the first step, the auto-encoder parameters θ_{e} and θ_{d} are updated assuming fixed speaker ID parameters θ_{id}, and in the second step, the parameters θ_{id} are updated assuming fixed θ_{e} and θ_{d} obtained in the first step, i.e.,

$$(\hat{\theta}_{\text{e}}, \hat{\theta}_{\text{d}}) = \arg\min_{\theta_{\text{e}}, \theta_{\text{d}}} E(\theta_{\text{e}}, \theta_{\text{d}}, \hat{\theta}_{\text{id}}), \tag{5.4}$$

$$\hat{\theta}_{\text{id}} = \arg\max_{\theta_{\text{id}}} E(\hat{\theta}_{\text{e}}, \hat{\theta}_{\text{d}}, \theta_{\text{id}}). \tag{5.5}$$

For adversarial training, the gradient reversal layer (GRL) is used (with λ being the GRL parameter) (Ganin and Lempitsky, 2015; Ganin et al., 2016). GRL acts as an identity function in forward propagation and inverts the sign of the loss function in backpropagation. Each parameter set is updated using the ADAM optimizer (Kingma and Ba, 2015). While all training speakers (healthy and pathological) are used for optimizing the reconstruction loss L_{ae}, we consider data only from healthy speakers to optimize the speaker ID loss L_{id}. This ensures that only non-pathological speaker variabilities are suppressed from the bottleneck representation.

Pathological speech discriminative representation

To learn pathological speech discriminative representations, the bottleneck representation of the auto-encoder is connected to a pathological speech classifier module. The same architecture of fully connected layers as for the speaker ID module (described in the previous subsection) is used for the pathological speech classifier module. However, differently from the speaker ID module, the final layer for the pathological speech classifier module consists of 2 output units since we are dealing with binary classification (i.e., pathological speech vs. typical speech). The parameters of this module are denoted by θ_{pc}.

The optimal parameters θ_{e}, θ_{d}, and θ_{pc} are computed as the ones simultaneously minimizing the auto-encoder reconstruction loss L_{ae} and the pathological speech detection loss L_{pc}, i.e.,

$$(\hat{\theta}_{\mathrm{e}}, \hat{\theta}_{\mathrm{d}}, \hat{\theta}_{\mathrm{pc}}) = \arg \min_{\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}} E(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}), \tag{5.6}$$

with

$$E(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}) = L_{\mathrm{ae}}(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}) + \alpha L_{\mathrm{pc}}(\theta_{\mathrm{e}}, \theta_{\mathrm{pc}}), \tag{5.7}$$

where α is the trade-off parameter between the two loss functions (cf. Section 5.3.3). Similarly to before, the ADAM optimizer is used for finding the optimal parameters.

Fusion of pathological speech discriminative representation and speaker identity-invariant representation with adversarial training

To jointly learn a speaker identity-invariant and pathological speech discriminative representation, we also consider training the auto-encoder using both auxiliary modules (described in previous subsections) through the optimization objective

$$(\hat{\theta}_{\mathrm{e}}, \hat{\theta}_{\mathrm{d}}, \hat{\theta}_{\mathrm{pc}}, \hat{\theta}_{\mathrm{id}}) = \arg \min_{\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}} \arg \max_{\theta_{\mathrm{id}}} E(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}, \theta_{\mathrm{id}}), \tag{5.8}$$

where

$$\begin{aligned} E(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}, \theta_{\mathrm{id}}) = {} & L_{\mathrm{ae}}(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}) \\ & + \alpha L_{\mathrm{pc}}(\theta_{\mathrm{e}}, \theta_{\mathrm{pc}}) - \lambda L_{\mathrm{id}}(\theta_{\mathrm{e}}, \theta_{\mathrm{id}}). \end{aligned} \tag{5.9}$$

The solution to (5.8) is approximated using a similar alternating training procedure as in the previously described adversarial training.

Pathological speech classification

After obtaining the bottleneck representation following any of the training procedures outlined in previous subsections, this representation is used to train a pathological speech classifier. The classifier architecture is identical to the auxiliary classifier module used before for training the pathological speech discriminative representation. The final decision for an unseen (test) speaker is made by applying soft voting on the classifier prediction scores for all input TF representations belonging to that speaker.

5.3.2 Proposed supervised dual representation learning (feature separation)

To obtain a bottleneck representation that contains less cues about speaker identities without using adversarial training, we propose to use a feature separation framework with two encoders as illustrated in Figure 5.4. In this framework, chunks of speech TF representations are

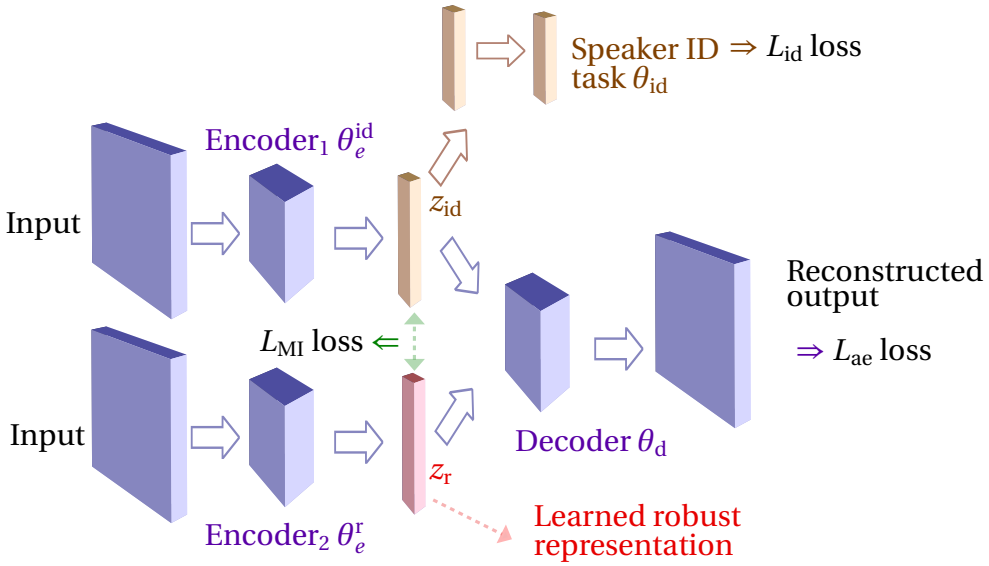


Figure 5.4 – Proposed feature separation framework to obtain speaker-invariant representation for pathological speech detection without using adversarial training. This framework includes two encoders, a single decoder, an auxiliary speaker ID module, and a MI minimizer. Two separate bottleneck representations are learned such that one is intended to encode only speaker identity cues guided by the auxiliary speaker ID task, and the other one is intended to contain less speaker identity information by minimizing the MI between the two representations.

projected into a pair of bottleneck representations using two encoders. A decoder is also trained to generate a reconstructed version of the input using the concatenated bottleneck representations to avoid loss of information embedded within both representations. The first bottleneck representation, denoted as z_{id}, is intended to contain only information about speaker identities, while the second (residual) bottleneck representation, denoted as z_{r}, is intended to contain no information about speaker identities. Therefore, the z_{id} representation is directly supervised by a speaker ID classifier while the z_{r} representation is isolated from the z_{id} representation by minimizing an MI criterion between z_{id} and z_{r}. Since all the modules, i.e., the two encoders, the decoder, and the auxiliary speaker ID classifier are jointly trained, it can be expected that the residual z_{r} bottleneck representation encodes speaker identity-invariant cues, making z_{r} a more robust representation for pathological speech detection.

The architecture of both encoders is the same and identical to the encoder module of the auto-encoder used in the previous framework (cf. Section 5.3.1). For the speaker ID classifier, the same architecture as in the previous framework is used. The decoder architecture is also

the same as the previous framework with the difference being the size of the first layer of the decoder. In the current feature separation framework, two bottleneck representations are concatenated before being fed to the decoder, hence the number of nodes in the first fully connected layer of the decoder would be twice the number of nodes in the first fully connected layer of the decoder used in the previous framework (cf. Section 5.3.1). The parameters of the two encoders generating bottleneck representations z_{id} and z_{r} are denoted by θ_e^{id} and θ_e^{r}, respectively, and similarly to before, the parameters of the decoder and speaker ID modules are denoted by, θ_{d} and θ_{id} respectively. This framework also consists of an MI estimator module which is needed for estimating and minimizing the MI criterion between z_{id} and z_{r} as will be explained in the following.

MI minimizer

To reduce the dependency between the two bottleneck representations, we consider minimizing MI between z_{id} and z_{r}. Considering z_{id} and z_{r} as two continuous random variables, the MI, $I(z_{\text{id}}, z_{\text{r}})$, is defined as the Kullback-Leibler (KL) divergence between the joint distribution and the product of marginal distributions of the two variables, i.e.,

$$I(z_{\text{id}}, z_{\text{r}}) = D_{KL}\big(p(z_{\text{id}}, z_{\text{r}}) || p(z_{\text{id}})p(z_{\text{r}})\big). \tag{5.10}$$

Since the MI computation is challenging for high-dimensional variables with unknown probability distributions, variational contrastive log-ratio upper bound (vCLUB) is proposed in Cheng et al. (2020) to calculate an upper bound for MI, i.e.,

$$I_{\text{vCLUB}}(z_{\text{id}}, z_{\text{r}}) = \mathbb{E}_{p(z_{\text{id}}, z_{\text{r}})}\left[\log q_\phi(z_{\text{id}}|z_{\text{r}})\right] - \mathbb{E}_{p(z_{\text{id}})}\mathbb{E}_{p(z_{\text{r}})}\left[\log q_\phi(z_{\text{id}}|z_{\text{r}})\right], \tag{5.11}$$

where $q_\phi(z_{\text{id}}|z_{\text{r}})$ is the variational approximation of $p(z_{\text{id}}|z_{\text{r}})$ which is parameterized in a Gaussian family $q_\phi(z_{\text{id}}|z_{\text{r}}) = \mathcal{N}(z_{\text{id}}|\mu(z_{\text{r}}), \sigma^2(z_{\text{r}})I)$ with mean $\mu(z_{\text{r}})$ and variance $\sigma^2(z_{\text{r}})$ being estimated by (MI estimator) neural networks with overall parameters of ϕ (Cheng et al., 2020). The MI estimator for the mean or variance is parameterized by a fully connected layer with 64 hidden units following a ReLU activation function that outputs a 128 dimensional vector representing $\mu(z_{\text{r}})$ and $\sigma^2(z_{\text{r}})$. For the network estimating the variance, a Tanh (hyperbolic tangent) activation function is also applied after the output. Due to identical encoder architectures used here (and used in the previous approaches to learn a single representation of dimension 128 (cf. Section 5.3.1)), the dimension of both bottleneck representations (z_{id} and z_{r}), and therefore the dimension of input and output of MI estimator networks, are also 128. The parameters of the MI estimator are approximated by maximizing the log-likelihood loss, i.e., $L_{\text{ll}}(\phi) = \log q_\phi(z_{\text{id}}|z_{\text{r}})$ as in Cheng et al. (2020). After obtaining the parameters of the MI estimator, we use vCLUB as our MI objective to be minimized, i.e., $L_{\text{MI}}(\theta_e^{\text{id}}, \theta_e^{\text{r}}) = I_{\text{vCLUB}}(z_{\text{id}}, z_{\text{r}})$.

Feature separation

Obtaining a speaker-invariant representation z_{r} is achieved through the optimization of the following objective function, where due to the presence of MI loss and MI estimators, optimal parameters are approximated using an alternating training procedure, i.e.,

$$(\hat{\theta}_e^{\mathrm{id}}, \hat{\theta}_e^{\mathrm{r}}, \hat{\theta}_{\mathrm{d}}, \hat{\theta}_{\mathrm{id}}) = \arg \min_{\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \theta_{\mathrm{d}}, \theta_{\mathrm{id}}} E(\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \theta_{\mathrm{d}}, \theta_{\mathrm{id}}, \hat{\phi}) \tag{5.12}$$

$$\hat{\phi} = \arg \min_{\phi} -L_{\mathrm{ll}}(\phi, \hat{\theta}_e^{\mathrm{id}}, \hat{\theta}_e^{\mathrm{r}}) \tag{5.13}$$

with

$$E(\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \theta_{\mathrm{d}}, \theta_{\mathrm{id}}, \hat{\phi}) = L_{\mathrm{ae}}(\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \theta_{\mathrm{d}}) + \lambda L_{\mathrm{id}}(\theta_e^{\mathrm{id}}, \theta_{\mathrm{id}}) + \beta L_{\mathrm{MI}}(\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \hat{\phi}), \tag{5.14}$$

where λ and β are the trade-off parameters for speaker ID and MI loss functions (cf. Section 5.3.3). All training speakers (healthy and pathological) are used for optimizing the reconstruction loss L_{ae} and L_{MI}, while similarly to before, we consider only data from healthy speakers to optimize the speaker ID loss L_{id}. This ensures that only non-pathological speaker variabilities are being pushed away from the residual bottleneck representation z_{r}.

Pathological speech classification

The final learned encoder generating the bottleneck representation z_{r} that contains less speaker identity cues is then used as input for pathological speech detection. The architecture of the classifier used for pathological speech detection is identical to the final and auxiliary classifier modules in Section 5.3.1. The final decision for an unseen (test) speaker is made by applying soft voting on the classifier prediction scores for all input TF representations belonging to that speaker.

5.3.3 Experimental Results

In this section, we first evaluate our proposed supervised single representation learning approaches. Then, we evaluate our proposed supervised dual representation learning approach based on the feature separation framework. The performance of the proposed approaches is compared to their corresponding baseline systems.

Baseline networks for supervised single representation learning

We consider two sets of baseline systems for the supervised single representation learning framework.

Without representation learning. In order to investigate the advantages of representation

learning through using a decoder in our supervised representation learning frameworks, we consider a baseline classifier trained in a fully supervised manner without including a decoder, and therefore, no reconstruction error loss is involved in the training procedure. Hence, there is no explicit attempt to learn a speech representation reconstructing the input. The classifier architecture is composed of an encoder (identical to the encoder module in Section 5.3.1) followed by a pathological speech classifier (identical to the classifier module in Section 5.3.1). We denote this system as Baseline$_0$.

With representation learning. To demonstrate the advantages of the obtained speaker identity-invariant and pathological speech discriminative representations, we consider the vanilla auto-encoder in (Vasquez-Correa et al., 2020) as the baseline system where the single bottleneck representation is learned using an auto-encoder (with the same architecture as in Section 5.3.1) without any supervision. As a comparison to the vanilla auto-encoder, we also consider the PCA dimensionality reduction method which transforms chunks of speech TF representations into a lower-dimensional vector (e.g., of size 128), while maintaining as much information as possible. Unlike auto-encoders, PCA uses a linear orthogonal transformation. We denote PCA-based unsupervised representation learning as Baseline$_1$ while vanilla auto-encoder-based unsupervised representation learning is denoted as Baseline$_2$. Furthermore, to investigate the suitability of supervised representation learning for suppressing irrelevant speaker identity information, we also train a speaker ID module on each of the learned representations. The architecture of this module is identical to the auxiliary speaker ID module in Section 5.3.1.

Baseline networks for supervised dual representation learning (feature separation)

For the supervised dual representation learning framework, we consider the following baseline systems. To investigate the effects of the auxiliary modules on feature separation, our baseline systems are based on excluding the supervision of these modules. Without using the speaker ID and MI minimizer modules in training, i.e., setting $\lambda = \beta = 0$ in (5.14), we obtain one baseline system that is trained without any supervision (unsupervised dual representation learning). Keeping the speaker ID module, but removing the MI minimizer during training, i.e., setting $\beta = 0$ in (5.14), we obtain a second baseline system.

Training and evaluation

As in Vasquez-Correa et al. (2020), the input TF representations for our frameworks, are Mel-scale representations of 500 ms segments of speech with 50% overlap. Mel-scale representations are computed using 32 ms Hamming windows with a frame shift of 4 ms and 126 Mel bands. Z-score normalization is applied to all input representations. Similarly to the previously proposed approach, i.e., pairwise distance-based CNNs, AP representations can also be used as inputs in this framework. Obtaining AP representations requires training phoneme-to-articulatory feature mappings using a large number of healthy speech recordings (cf. Section 2.4.5). To avoid AP training in this framework, we choose Mel-scale representations

as have been previously used for unsupervised representation learning for pathological speech detection (Vasquez-Correa et al., 2020).

For training and evaluation, we use the same validation strategy for our databases as before, i.e., a stratified speaker-independent 10-fold and 5-fold cross-validation on the PC-GITA and MoSpeeDi databases, respectively.

In each training fold, a development fold of the same size as the test fold is set aside for early-stopping in training the final pathological speech classifier. For the speaker ID auxiliary task, utterances from the healthy speakers in the training set (i.e., 45 speakers for PC-GITA and 16 speakers for MoSpeeDi) are split without overlap into 50% train, 25% development, and 25% test sets. Cross-entropy is used for the auxiliary loss functions L_{id} and L_{pc}, and for the final pathological speech classification in all systems, whereas mean square error (MSE) of the reconstruction is used for the auto-encoder loss L_{ae}.

Since our supervised representation learning framework is composed of different modules for different tasks, our preliminary results showed that the learning rate for each module should be different, e.g., for the speaker identity classifier, a higher learning rate is required compared to the pathological speech classifier. Setting different learning rates is particularly important when adversarial training is used. After we set an initial learning rate of 10^{-5} for the auto-encoder modules in both frameworks, we set the learning rate for pathological speech classifier and speaker ID classifier modules to 10 and 10^2 times higher, respectively. All models are trained with a batch size of 128. When using the pathological speech classifier module in the single representation learning framework, early-stopping is performed where the learning rate for the overall model is halved each time the auxiliary pathological speech detection loss on the development set does not decrease for 5 consecutive iterations. Training is stopped either after 50 epochs or after the auto-encoder learning rate has decreased beyond 10^{-6}. In the dual representation learning framework, we also trained the systems for 50 epochs. For the final pathological speech detection after representation learning in all systems, the pathological speech classifier is trained with a learning rate 10^{-4} while freezing the encoder parameters. The learning rate for the classifier is halved each time the classification loss on the development set does not decrease for 5 consecutive iterations. Training is stopped either after 50 epochs or after the classifier learning rate has decreased beyond 10^{-5}.

The pathological speech detection performance is evaluated in terms of accuracy (i.e., percentage of correctly classified healthy and pathological speakers) and the AUC. The performance for the speaker ID task is evaluated for unseen (test) utterances also using accuracy (i.e., percentage of correctly identified speakers) and AUC. To reduce the impact of the random seed on the final model parameters all networks are trained with 5 different random seeds. The reported performance measures are the mean and standard deviation of the performance obtained by models trained using different seeds.

To select the hyper-parameters λ and α of the proposed single representation learning framework (cf. (5.3) and (5.7)), we use grid-search for the set of values $\lambda \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$

and $\alpha \in \{0.0005, 0.0001, 0.001, 0.01, 0.1, 0.5\}$. The final hyper-parameters λ and α are selected as the ones yielding the highest mean pathological speech detection accuracy on the development set. For the fusion approach in Section 5.3.1, the hyper-parameters used in (5.9) are optimized using a grid search for the set of previously mentioned values for λ and α. For feature separation framework, the hyper-parameters λ and β (cf. (5.14)) are selected using grid-search for the set of values $\lambda \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$ and $\beta \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$. The final hyper-parameters λ and β are selected as the ones yielding the highest mean pathological speech detection accuracy on the development set.

Results for supervised single representation learning

In this section, we first present the results obtained for the MoSpeeDi database using the supervised single representation learning, where we show that such representation learning is not suitable for this database. Then, we present results using the larger PC-GITA database.

Table 5.6 presents the pathological speech detection accuracy and AUC values obtained using the baseline systems and the supervised single representations learning frameworks on the MoSpeeDi database. As it can be observed, the highest performance is obtained using Baseline_0 where no representation learning is used, whereas the representation learning frameworks have failed to further improve the performance. It can be seen that pathological speech discriminative training improved the performance in comparison to the unsupervised representation learnings using PCA (Baseline_1) and vanilla auto-encoder (Baseline_2), while the adversarial speaker invariant training yielded the lowest performance. It should be noted that the architecture used in our representation learning frameworks are adapted from Vasquez-Correa et al. (2020), where these architectures were optimized for the PC-GITA database. These architectures have not been further optimized for the MoSpeeDi database, which is a significantly smaller database than the PC-GITA database. The reason that representation learning is not as effective on the MoSpeeDi database might be due to non-suitable training parameters and architectures. Furthermore, the reason that speaker-invariant representation learning yields the lowest performance can be attributed to the limited number of healthy speakers in the training set (i.e., 16) required for training the speaker ID module.

Considering the supervised single representation learning framework using the PC-GITA database, Table 5.7 presents the pathological speech detection accuracy and AUC values obtained using the baseline system without exploiting any representation learning (i.e., Baseline_0), the baseline representation using PCA (i.e., Baseline_1), the vanilla auto-encoder from Vasquez-Correa et al. (2020) learned without any supervision (Baseline_2)[1], and the proposed supervised representations learned through auxiliary tasks.

[1] It should be noted that the auto-encoder used in Vasquez-Correa et al. (2020) was trained on a larger healthy speech database. However, although not presented here, using the same healthy speech database for training the auto-encoder did not result in a better performance than the performance obtained using only the PC-GITA database.

Table 5.6 – Mean and standard deviation of the speech pathology detection accuracy [%] and AUC score using the Baseline_0, unsupervised single representation learning (i.e., Baseline_1 and Baseline_2) and the proposed supervised single representation learning approaches on the French MoSpeeDi database.

No representation learning		
System	Accuracy (%)	AUC
Baseline_0 (without a decoder)	**73.5** $\pm$ 2.5	**0.89** $\pm$ 0.03
Representation learning		
Auxiliary task in representation learning	Accuracy (%)	AUC
No auxiliary task (Baseline_1, i.e., PCA)	62.0 $\pm$ 1.9	0.70 $\pm$ 0.02
No auxiliary task (Baseline_2)	68.0 $\pm$ 2.9	0.73 $\pm$ 0.02
Adversarial speaker invariant training	52.0 $\pm$ 1.9	0.68 $\pm$ 0.03
Pathological speech discriminative training	70.0 $\pm$ 2.7	0.81 $\pm$ 0.02
Fusion (speaker invariant+pathological speech discriminative training)	55.0 $\pm$ 2.2	0.73 $\pm$ 0.05

Table 5.7 – Mean and standard deviation of the speech pathology detection accuracy [%] and AUC score using the Baseline_0, unsupervised single representation learning (i.e., Baseline_1 and Baseline_2) and the proposed supervised single representation learning approaches on the Spanish PC-GITA database.

No representation learning		
System	Accuracy (%)	AUC
Baseline_0 (without a decoder)	71.8 $\pm$ 3.0	0.80 $\pm$ 0.02
Representation learning		
Auxiliary task in representation learning	Accuracy (%)	AUC
No auxiliary task (Baseline_1, i.e., PCA)	53.8 $\pm$ 2.0	0.61 $\pm$ 0.03
No auxiliary task (Baseline_2)	60.8 $\pm$ 1.7	0.72 $\pm$ 0.02
Adversarial speaker invariant training	77.0 $\pm$ 4.2	0.85 $\pm$ 0.03
Pathological speech discriminative training	73.0 $\pm$ 3.2	**0.87** $\pm$ 0.02
Fusion (speaker invariant+pathological speech discriminative training)	**78.0** $\pm$ 1.9	**0.87** $\pm$ 0.02

It can be observed that using the representations learned by any of the proposed auxiliary tasks improves the performance of pathological speech detection compared to using the baseline unsupervised learned representations by either PCA or vanilla auto-encoder. Furthermore, both supervised representation learnings have improved the performance in terms of both measures when there is no explicit representation learning is involved in training (i.e., Baseline_0). When comparing the two proposed supervised representation learning approaches, a larger performance improvement is observed in terms of accuracy for the speaker-invariant training,

while a larger performance in terms of AUC is achieved by pathological speech discriminative training. Furthermore, fusing both auxiliary tasks to obtain a robust and discriminative representation yields a better pathological speech detection accuracy than other representations, clearly outperforming the unsupervised baseline systems as well. It can be observed that while the fusion of auxiliary tasks improves the pathological speech detection accuracy as opposed to using any of the auxiliary tasks, the resulting AUC is still the same as when using pathological speech discriminative training.

In summary, the results presented in Table 5.7 confirm the advantages of supervised representation learning for pathological speech detection on the PC-GITA database. However, similar advantageous were not observed for the MoSpeeDi database. We suspected that is due to suboptimal training parameters and the small amount of training data in this database negatively affecting the considered frameworks. For this reason, only the PC-GITA database is considered for the results presented in the following.

To investigate the suppression of irrelevant speaker identity information in each of the supervised representations as opposed to the unsupervised representations, Table 5.8 presents the accuracy and AUC values obtained for the speaker ID task on all the different representations using the PC-GITA database.

It can be observed that using the baseline (unsupervised) representations learned by PCA results in the highest speaker ID performance, i.e., classifying healthy speakers' identities with almost 100% accuracy. The unsupervised representation learned by the vanilla auto-encoder gives lower speaker ID performance than PCA but nevertheless, a higher performance than the supervised learned representations is obtained. We observed that PCA yields representations resulting in lower MSE compared to the representations obtained through a vanilla auto-encoder trained with a limited number of epochs. Since in representations obtained by PCA, there is less information lost (i.e., a lower MSE), PCA results in the highest speaker ID performance. This result confirms that unsupervised training yields representations containing speaker identity cues, reducing as a result the generalization and the final performance

Table 5.8 – Mean and standard deviation of the speaker ID accuracy [%] and AUC score using the unsupervised single representation learning (i.e., Baseline_1 and Baseline_2) and the proposed supervised single representation learning approaches on the Spanish PC-GITA database.

Auxiliary task in representation learning	Accuracy (%)	AUC
No auxiliary task (Baseline_1, i.e., PCA)	99.1 ± 0.4	1.00 ± 0.00
No auxiliary task (Baseline_2)	56.3 ± 2.5	0.98 ± 0.00
Adversarial speaker invariant training	5.2 ± 2.2	0.67 ± 0.05
Pathological speech discriminative training	35.2 ± 4.4	0.94 ± 0.00
Fusion (speaker invariant+pathological speech discriminative training)	14.0 ± 5.4	0.80 ± 0.07

of pathological speech detection (cf. Table 5.7). Further, as expected, the lowest speaker ID performance is observed for the speaker identity-invariant representations obtained using adversarial training. These results confirm the suitability of adversarial training to reduce the presence of irrelevant speaker identity cues in the bottleneck representation. Finally, it can be observed that although the pathological speech discriminative feature representation results in a higher speaker ID performance than adversarial training, it still yields a lower speaker ID performance than the unsupervised baseline representations. This result shows that supervising the auto-encoder training such that a discriminative feature representation for pathological speech detection is learned inherently reduces the presence of speaker identity cues, since they are irrelevant to the pathological speech detection task. For the fusion of auxiliary tasks, we observe lower speaker ID performance than pathological speech discriminative feature representation due to the presence of adversarial training.

It should be noted that we also investigated the presence of other pathology-unrelated cues (e.g., age and gender) in the learned representations using the PC-GITA database. We obtained gender-invariant and age-invariant representations with similar adversarial training as in Section 5.3.1. In our further analysis (results that we have omitted here), we observed that suppressing age or gender cues does not improve the pathological speech detection performance compared to suppressing speaker identity cues. Our analysis showed that age cues are not captured in any of the learned representations obtained by unsupervised and supervised auto-encoders (yielding accuracy $< 50\%$ and AUC < 0.56), therefore removing age cues using adversarial training did not improve the performance. Surprisingly, we observed that gender cues are present in all learned representations, with more cues present in the pathological speech discriminative representation (yielding accuracy $> 90\%$ and AUC > 0.98). This suggests that gender cues might be useful for pathology detection as the speech disorder can affect people of different genders differently.

Comparing the distance-based CNN and single representation learning

Before comparing the performance of our supervised representation learning to our previously proposed pairwise distance-based CNN (cf. Section 5.2) on the PC-GITA database, we should first point out the inevitable differences in the used speech material for training both systems. Pairwise distance-based CNN relies on using a limited number (< 25) of word utterances with the same phonetic content from both healthy and pathological speakers. Supervised representation learning approaches do not require any phonetic constraints in the speech material, however more speech data (resulted from uttering sentences, text, and words by each speaker) are used in their training to achieve a robust performance. Comparing the results obtained by our supervised representation learning in Table 5.7 to the results obtained by pairwise distance-based CNN (cf. Table 5.5) while ignoring these inevitable differences between the two approaches, it can be observed that their performance is not largely different.

Table 5.9 – Mean and standard deviation of the speech pathology detection accuracy [%] and AUC score obtained from learned bottleneck representations z_r and z_{id} using dual representation learning approaches on the Spanish PC-GITA database.

Speaker ID task	MI minimizer	representations z_r		representation z_{id}	
		Accuracy (%)	AUC	Accuracy (%)	AUC
Baseline					
×	×	57.2 ± 5.4	0.72 ± 0.02	$\mathbf{56.2} \pm 3.8$	$\mathbf{0.72} \pm 0.03$
✓	×	61.4 ± 3.4	0.75 ± 0.02	55.6 ± 4.0	0.67 ± 0.02
Proposed					
✓	✓	$\mathbf{75.2} \pm 3.5$	$\mathbf{0.82} \pm 0.03$	54.2 ± 6.2	0.63 ± 0.03

Results for supervised dual representation learning (feature separation)

In this section, the feature separation framework to obtain a representation robust to speaker identity without adversarial training is evaluated by considering the PC-GITA database. As mentioned before, if the feature separation is successful, the bottleneck representation z_{id} is expected to contain more speaker identity cues, whereas the bottleneck representation z_r is expected to contain less speaker identity cues. Therefore, it is expected that using z_r for pathological speech detection task performs better than using z_{id}.

Table 5.9 presents the pathological speech detection accuracy and AUC values using the two learned representations trained by the baseline frameworks and by our proposed supervised feature separation frameworks. Based on the results presented in this table, several observations can be made.

- First, using either of the representations obtained from the baseline framework (i.e., z_{id} or z_r) where neither speaker ID nor MI minimizer is included in the training, a similarly low pathological speech detection performance is achieved. This is to be expected as the two representations are trained similarly in the unsupervised framework.
- Second, including only the speaker ID module in the training slightly improves the pathological speech detection performance of z_r compared to the unsupervised baseline system, while a slight performance decrease can be observed for z_{id}. This is to be expected due to the direct supervision of the representation z_{id} to include more speaker identity cues, and hence, decreasing the performance of pathological speech detection. However, in both baseline systems, no supervision is used for isolating speaker identity cues from z_r, therefore their performance when using z_r is comparable.
- Third, the representation z_r learned by the proposed method gives the highest performance outperforming the baseline systems, while the performance using the representation z_{id} remains low. Considering the two representations obtained by our proposed

Table 5.10 – Mean and standard deviation of speaker ID accuracy [%] and AUC score obtained from learned bottleneck representations z_r and z_{id} using dual representation learning approaches on the Spanish PC-GITA database.

Speaker ID task	MI minimizer	representation z_r		representation z_{id}	
		Accuracy (%)	AUC	Accuracy (%)	AUC
Baseline					
×	×	58.3 ± 2.1	0.98 ± 0.00	56.1 ± 4.5	0.98 ± 0.00
✓	×	49.6 ± 3.2	0.98 ± 0.00	87.0 ± 2.8	1.00 ± 0.00
Proposed					
✓	✓	5.0 ± 5.2	0.67 ± 0.09	71.7 ± 4.3	0.99 ± 0.00

feature separation framework using both speaker ID and MI minimizer modules, these results confirm that the z_{id} contains non-relevant information for pathological speech detection while z_r is the most informative representation for the task. This can be attributed to the efficacy of the performed separation of speaker identity cues by our framework. Furthermore, by comparing the results obtained using z_r trained by the proposed method in Table 5.9 to the adversarial speaker invariant training results using the single representation learning in Table 5.7, it can be observed that the pathological speech detection performance using the feature separation framework without adversarial training is not largely different than using single representation learning with adversarial training.

To further analyze the performance of the feature separation framework, we investigate the presence of speaker identity cues in all representations by evaluating the performance of a speaker ID classifier trained on the obtained representations. Table 5.10 presents the accuracy and AUC values obtained for the speaker ID task using the two learned representations trained by the baseline frameworks and by our proposed supervised feature separation framework. It can be observed that using representations learned by the unsupervised baseline system (i.e., z_{id} or z_r) where neither speaker ID nor MI minimizer is included in the training, similar speaker ID performance is achieved. This confirms that the two encoders generate feature representations containing a similar amount of speaker identity cues. In addition, it can be observed that including only the speaker ID module in the training only decreases the speaker ID performance using the representation z_r in terms of accuracy compared to the unsupervised baseline system, while, as expected, a significant increase is observed when using the representation z_{id} due to the direct supervision of z_{id} to include more speaker identity cues. Finally, it can be observed that the representation z_r learned by the proposed method gives the lowest speaker ID performance, while the performance using the representation z_{id} is higher than the performance of unsupervised baseline system. These results confirm the applicability of the proposed feature separation by our framework in which speaker-specific

cues during training are being isolated from z_r to obtain a more robust representation for pathological speech detection.

5.4 Summary

As an alternative to using handcrafted acoustic features in the proposed T-GDA approach in Chapter 4 which also requires time-alignment, in this chapter, we have proposed two CNN-based frameworks to learn and exploit high-level representations for automatic pathological speech detection. We were motivated by the necessity of alleviating overfitting issues dictated by the small size of the typically available databases and also by the necessity of learning more robust and relevant features for the pathological speech detection task.

In the first approach, we have proposed analyzing pairwise distance matrices where we represent utterances through articulatory posteriors and consider pairs of phonetically-balanced representations, with one representation from a healthy speaker (i.e., the reference representation) and the other representation from the test speaker (i.e., test representation). This approach benefits from pairwise training, which inherently guides the network to extract more robust features (robust to irrelevant speaker-specific information) by considering many paired representations from different speakers. Given such pairs of reference and test representations, features are first extracted using a feature extraction front-end, a frame-level distance matrix is computed, and the obtained distance matrix is considered as an image by a CNN-based binary classifier. The feature extraction, distance matrix computation, and CNN-based classifier are jointly optimized in an end-to-end framework. Experimental results on the considered Spanish and French databases of healthy and pathological speakers have shown that the proposed approach yields a high pathological speech detection performance, outperforming other CNN-based baseline approaches. However, such a system relies on using utterances with the same phonetic content from both healthy and pathological speakers.

In the second approach, relaxing the phonetic constraints required for the first approach, we have focused on explicitly learning a high-level abstract representation from short segments of speech that is robust to pathology-unrelated cues such as speaker identity information and/or is discriminative for pathology detection. To this end, we have exploited supervised auto-encoders to extract robust and discriminative speech representations for speech pathology detection. To reduce the influence of speaker variabilities unrelated to pathology, we have proposed two approaches to obtain speaker identity-invariant representations: i) we have trained an auto-encoder jointly with adversarial training a speaker ID module resulting in a single representation and ii) we have used a non-adversarial feature separation framework by training a dual encoder and a single decoder generating two representations. To enforce speaker identity cues to be present only in one of the representations, we have supervised one of the representations with a speaker ID task while minimizing a MI criterion between the two representations. In addition, to obtain a discriminative representation, we have proposed to jointly train an auto-encoder and a pathological speech classifier. Experimental results

on the Spanish database have shown that the proposed supervised single representation learning frameworks yield more robust and discriminative representations for automatically classifying pathological speech, outperforming the baseline system without explicit representation learning and also the baseline systems with unsupervised representation learning. However, the proposed representation learning frameworks did not generalize to the smaller French database. Furthermore, evaluating the feature separation learning framework on the Spanish database has confirmed the success of such a framework to obtain speaker-invariant representations without adversarial training.

6 Pathological speech intelligibility assessment based on a short-time objective intelligibility measure

In this chapter we propose a measure to assess speech intelligibility of pathological speech based on the extended short-time objective intelligibility assessment. This measure requires creating utterance-dependent reference representations from speech signals of multiple healthy (fully intelligible) speakers perfectly matching the phonetic content of the pathological speech signal. To increase its flexibility, we also propose to use synthetic speech generated by text-to-speech systems to create reference representations. The applicability of the proposed intelligibility measure in this chapter is experimentally investigated across databases and compared to many state-of-the-art pathological speech intelligibility measures.

6.1 Introduction

In Chapter 3, we introduced state-of-the-art approaches for automatic pathological speech intelligibility assessment which were broadly categorized into blind approaches and non-blind approaches. Blind approaches often underperform since they do not use fully intelligible data from healthy speakers as references to better estimate speech intelligibility. Non-blind approaches however tend to combine a large number of features for intelligibility prediction, which increases the risk of over-fitting and limits the performance in unseen data (due to the lack of a large amount of pathological training data with available subjective intelligibility scores). More successful non-blind approaches, e.g., ASR or GMM-based approaches (cf. Chapter 3), are typically complex and require a large number of healthy speech recordings for training, which might not be feasible for low-resource languages.

To tackle the drawbacks of state-of-the-art techniques, in this chapter we propose a non-blind pathological speech intelligibility measure based on the extended short-time objective intelligibility (P-ESTOI). P-ESTOI is motivated by the extended STOI (ESTOI) measure which is an objective intelligibility measure commonly used in speech enhancement. ESTOI has been successful in estimating the intelligibility of speech contaminated by temporally modulated noise (Jensen and Taal, 2016). Direct application of enhancement objective measures to assess

pathological speech is difficult since they are typically based on comparing time-aligned noisy and reference (clean) signals. While the pathological speech signal can be viewed as a noisy signal, the reference signal, i.e., the non-impaired and fully intelligible version of the patient's speech signal, is clearly not available. Hence, we propose to use a temporal clustering method based on DTW to create an utterance-dependent reference representation from multiple healthy speakers. Intelligibility is then assessed through time-alignment of the pathological utterance with the utterance-dependent reference representation using DTW and computing the short-time spectral correlation between the two aligned representations. P-ESTOI takes speech perception and distortion into account, and unlike state-of-the-art measures, has a simple structure, minimizes the risk of overfitting by providing a single feature as an intelligibility score instead of relying on a large number of features, and does not require any training or a large number of healthy speech recordings.

For assessing the intelligibility of a sample utterance from a patient in P-ESTOI, recordings of the same utterance from several healthy speakers are needed. Consequently, P-ESTOI cannot be used in scenarios where healthy recordings perfectly matching the phonetic content of the pathological speech signal are not available. To deal with these scenarios, we also propose to exploit synthetic speech generated by text-to-speech (TTS) systems to create intelligible reference models in P-ESTOI. This way, P-ESTOI becomes a flexible measure which can also be used in phonetically-unbalanced scenarios (i.e., in scenarios where recordings from several healthy speakers uttering the same utterances as the pathological speaker are not available). This idea is motivated by the substantial progress made in the TTS field to generate high-quality synthesized speech capturing characteristics of intelligible natural speakers (Hinterleitner et al., 2013). Using TTS systems as an "average" intelligible speaker has already been successfully exploited in the past for different applications. For example, in Anumanchipalli et al. (2012), synthetic speech is used for voice disorder detection by extracting acoustic features characterizing the deviation of the test speech signal from its synthesized counterpart. In Soldo et al. (2012), TTS systems are used to generate reference templates in template-based ASR systems, showing comparable ASR performance to generating reference templates using natural speech. To our knowledge, the suitability of synthetic speech references for pathological speech intelligibility assessment has never been investigated.

The rest of the chapter is organized as follows. Section 6.2.1 presents an overview of the ESTOI measure in speech enhancement. Section 6.2.2 provides details on our proposed P-ESTOI intelligibility measure with natural healthy speech used to create fully intelligible references to assess the speech intelligibility of pathological speakers. Section 6.2.3 presents details on extending P-ESTOI measure by using synthetic speech to create intelligible references. The experimental results and discussions for two considered scenarios are described in Section 6.3, where P-ESTOI measure is compared to state-of-the-art measures, its robustness to non-pathological variations, e.g., gender and age is empirically analyzed, and its performance using synthetic speech references as opposed to natural speech references is investigated. Finally, Section 6.4 summarizes this chapter.

6.2 Proposed approach

In this section, our proposed intelligibility measure to automatically assess the intelligibility of utterances from patients is described. Since our measure is based on the ESTOI measure in speech enhancement, we first give an overview of the ESTOI measure. Then, our proposed pathological speech intelligibility measure using either healthy speech references or synthetic speech references is presented.

6.2.1 Overview of the extended short-time objective intelligibility measure

The ESTOI measure as defined in Jensen and Taal (2016) requires a clean and a degraded speech signal, which are assumed to be time-aligned. To estimate speech intelligibility of a sample degraded utterance, first, one-third octave band analysis is applied to the TF representation of both clean and degraded signals corresponding to that utterance, yielding in total J one-third octave bands. We denote the $J \times T$-dimensional time-aligned one-third octave band representations of the clean and degraded signals for the sample utterance n as $\mathbf{R^n}$ and $\mathbf{P^n}$, respectively, with T being the total number of frames. TF-units are denoted by $R^n(j,i)$ and $P^n(j,i)$, with j denoting the octave band index and i denoting the frame index.

To compute ESTOI, an intermediate intelligibility measure $b(t)$ is first computed from a region of I consecutive normalized TF-units, with $i \in \{t,(t+1),\ldots,(t+I-1)$ for $t \leq T-I+1$. All I consecutive TF-units in each band of the clean and degraded representations are mean and variance normalized. For each time frame, the linear correlation coefficient between J frequency bands is computed. Denoting by $\tilde{R}^n(j,i)$ and $\tilde{P}^n(j,i)$ the mean and variance normalized TF-units of each representation, $b(t)$ is computed as the average of the spectral linear correlation coefficients across I consecutive time frames (Jensen and Taal, 2016), i.e.,

$$b(t) = \frac{1}{I} \sum_{i=t}^{t+I-1} \frac{\sum_{j=1}^{J} \left(\tilde{R}^n(j,i) - \overline{\tilde{R}^n(j,i)}\right)\left(\tilde{P}^n(j,i) - \overline{\tilde{P}^n(j,i)}\right)}{\sqrt{\sum_{j=1}^{J} \left(\tilde{R}^n(j,i) - \overline{\tilde{R}^n(j,i)}\right)^2 \sum_{j=1}^{J} \left(\tilde{P}^n(j,i) - \overline{\tilde{P}^n(j,i)}\right)^2}}, \tag{6.1}$$

where $\overline{\tilde{R}^n(j,i)} = \frac{1}{J} \sum_{j=1}^{J} \tilde{R}^n(j,i)$ and $\overline{\tilde{P}^n(j,i)}$ is similarly defined. The intelligibility score of the degraded signal corresponding to the utterance n (denoted as IS^n) is finally computed as the average of the intermediate measure over all frames:

$$\mathrm{IS}^n = \frac{1}{(T-I+1)} \sum_{t} b(t). \tag{6.2}$$

It should be noted that ESTOI measure does not assume mutual independence between contributions of frequency bands to intelligibility unlike its STOI variant (Taal et al., 2010), where temporal correlations are analyzed instead of spectral ones.

It is worth mentioning that the intelligibility predictions given by IS^n in (6.2) should not

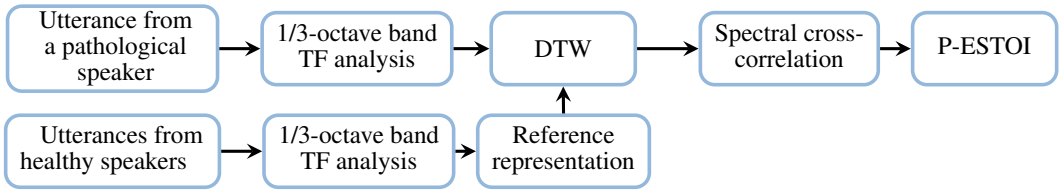


Figure 6.1 – Block diagram of the proposed pathological intelligibility measure P-ESTOI. The reference representation (template) is obtained from DTW-based clustering of healthy speakers' representations (after 1/3-octave band TF analysis). A test (possibly pathological) utterance is then compared by DTW to the reference representation to estimate spectral correlations, which are then used to calculate the P-ESTOI intelligibility measure (according to (6.2)).

be interpreted as an absolute intelligibility score (the percentage of words understood by listeners), but rather should be treated as an index, i.e., expected to be correlated with absolute subjective intelligibility scores. For learning the corresponding mapping IS^n to final absolute intelligibility scores, a separate test and train data including data from many pathological speakers with different levels of intelligibility is required. Such mapping will be dependent on the language and the speech material. Therefore, in this thesis, we did not attempt to learn such mappings.

6.2.2 Pathological intelligibility assessment using healthy speech references

Pathological speech intelligibility is a measure of the influence of the speech production deficit of a patient on a listener's perceptual understanding (Ansel and Kent, 1992), with pathological and healthy speech being differently perceived. We hypothesize that quantifying the divergence of a perceptual acoustic representation of pathological speech from healthy (intelligible) speech yields a reliable pathological intelligibility assessment technique. Therefore, we propose to use the simple perceptual acoustic representation used in ESTOI measure, i.e., the previously described one-third octave band representation in Section 2.4.2. By computing the spectral correlations between the octave band representations of a reference and time-aligned test signal, an estimate of speech intelligibility can be derived.

For pathological intelligibility assessment, the test signal is the pathological signal while a time-aligned (fully intelligible) reference signal is not available. Therefore, we propose to create an utterance-dependent reference representation from recordings from multiple healthy speakers using DTW. DTW is then also used to align the pathological speech representation to the reference representation before computing the intelligibility measure based on the spectral cross-correlation of the aligned representations. The block diagram of the resulting pathological speech intelligibility measure, i.e., P-ESTOI is illustrated in Figure 6.1. As depicted in this figure, the computation of P-ESTOI measure relies on i) creating an utterance-dependent (intelligible) reference representation, ii) aligning the considered pathological representation to the reference representation using DTW, and iii) computing the spectral

correlation between the two aligned representations to estimate the intelligibility.

In the following, the time alignment, the method proposed to create utterance-dependent reference signals from healthy speakers, and finally, the intelligibility assessment is described.

Time alignment

Let $\mathbf{X}_s^n$ denote the $J \times M$-dimensional octave band representation of the sample utterance n from speaker s, with M denoting the total number of time frames. Similarly, we define $\mathbf{X}_p^n$ to be the $J \times N$-dimensional octave band representation from another speaker p with N being the total number of time frames in $\mathbf{X}_p^n$. The representations $\mathbf{X}_s^n$ and $\mathbf{X}_p^n$ are typically not aligned (due to different speakers and speaking rates) and are generally of different lengths, i.e., $M \neq N$. These two representations are aligned through DTW, using a simple Euclidean distance as the cost function (Rabiner and Juang, 1993). DTW finds T-dimensional warping paths $\boldsymbol{\phi}_{s,p}$ and $\boldsymbol{\phi}_{p,s}$, with $T \geq \max[M, N]$, such that the warped representations $\mathbf{X}_s^n(\boldsymbol{\phi}_{s,p})$ and $\mathbf{X}_p^n(\boldsymbol{\phi}_{p,s})$ are point-to-point aligned sequences.

Utterance-dependent reference representations

For each considered utterance, a healthy speaker r is randomly selected, with $r \in \{1, \ldots, R\}$ and R being the total number of healthy speakers. Let us denote the one-third octave band representation of the utterance n from healthy speaker c by $\mathbf{H}_c^n$. Using DTW, the octave band utterance representation $\mathbf{H}_r^n$ is separately time-aligned with the representations from all remaining healthy speakers. For each frame in $\mathbf{H}_r^n$, we extract all frames mapped to it by DTW from the representations of all remaining speakers. The representation for each reference frame is then created by taking the mean of all extracted aligned frames. The final reference representation for the considered utterance denoted by $\mathbf{R}^n$ is then obtained by concatenating all reference frames so obtained. It should be noted that using such an approach results in a reference representation that has the same length as the initial randomly selected representation $\mathbf{H}_r^n$. Our experimental results suggest that the computed P-ESTOI measure is not sensitive to the selected initial reference representation.

Intelligibility assessment

To assess intelligibility, the one-third octave band representation for the considered test utterance is computed and aligned to the created reference template using DTW, with Euclidian distances as local scores. Due to different speaking rates, the aligned representations will obviously have repeated frames, i.e., after alignment, the shorter representation is likely to be expanded by repeating several frames. In Darley et al. (1969), it was shown that for diseases such as CP and ALS, the speaking rate did not show a high correlation with speech intelligibility. However, the repeated frames in the reference or patient representation will clearly affect the computed intelligibility measures. To discard the differences in speaking rates, these repeated

frames are removed before computing the P-ESTOI. Denoting the TF-units of the aligned healthy reference of the sample utterance n as $R^n(j,i)$ and pathological test representations of the the same utterance from the pathological speaker k (with repeated frames discarded) as $P^n(j,i)$, P-ESTOI, the intelligibility score of the sample utterance n from patient k (denoted here as IS_k^n) is computed using (6.2).

6.2.3 Pathological intelligibility assessment using synthetic speech references

As described in Section 6.2.2, to evaluate the intelligibility of an utterance from a pathological speaker, P-ESTOI creates a reference representation based on recordings of the same utterance from multiple healthy speakers. In practice, however, such recordings are not always available. To make P-ESTOI a flexible measure that can be used in scenarios where such recordings are not available, in this section we propose to generate the reference representation using synthetic utterances generated with high-quality state-of-the-art TTS systems.

We propose to use a Deep Neural Network (DNN)-based TTS system inspired by the Merlin TTS system (Wu et al., 2016). The Merlin TTS system has been used as a benchmark for assessing the quality of TTS systems in the *Blizzard Challenge* in 2016 (King and Karaiskos, 2016) and 2017 (King et al., 2017). It has been shown that such a system yields high-quality synthesized signals, outperforming systems based on Hidden Markov Models in terms of naturalness and intelligibility (Wu et al., 2016). We train this system on multiple healthy speakers. For each sample utterance n in the pathological speech signal, we generate multiple synthesized reference utterances. The reference representation $\mathbf{R}^n$ is then computed following the same procedure as in Section 6.2.2. However, instead of using healthy speech recordings of the same utterance, we use synthesized speech of the same utterance from multiple TTS systems trained on multiple healthy speakers. Although following such an approach requires multiple healthy speech recordings to train appropriate TTS systems, it does not require healthy recordings of exactly the same utterances that are present in the pathological speech signal.

6.3 Experimental results

In this section, first, the performance of the proposed P-ESTOI measure using healthy speech references, as well as its generalisation capabilities across languages and diseases is investigated on two considered databases, i.e., English-speaking CP patients database and Dutch-speaking patients with hearing impairment database (cf. Section 3.2). P-ESTOI measure is also compared to several state-of-the-art measures. In addition, the robustness of the P-ESTOI measure using healthy speech references to gender and age variations is empirically analyzed. Finally, the performance of P-ESTOI using synthetic speech references as opposed to natural speech references is extensively investigated for the English-speaking CP patients.

6.3.1 Algorithmic settings, evaluation, scenarios and state-of-the-art measures

To compute P-ESTOI, the TF analysis is performed using a 32 ms Hamming window with an overlap of 50% (cf. Section 2.4.2). The number of consecutive frames I considered for the correlation is 15 and the number of one-third octave bands J is 15.

For training the TTS systems required in P-ESTOI with synthetic speech references, we consider the CMU ARCTIC database consisting of recordings of 1132 phonetically-balanced utterances from 4 US English-speaking healthy speakers (2 males, 2 females) (Kominek and Black, 2004). To compute reference representations from synthetic speech signals, we train 4 TTS systems using these healthy recordings of the 4 healthy speakers in the CMU ARCTIC database. To this end, we use a DNN-based state-of-the-art Merlin TTS system in conjunction with the Festival front-end, two bidirectional long short-term memory networks as duration and acoustic models, and the WORLD vocoder. For details on the TTS systems and the training procedure, the reader is referred to Schnell and Garner (2018); Marelli et al. (2019). By training a TTS system for each speaker, we get 4 speaker-dependent TTS systems.

To evaluate all considered measures, we use the Pearson correlation coefficient (R) and the Spearman rank correlation coefficient (R_s) between the automatically estimated intelligibility (as the mean across all considered utterances) and the subjective intelligibility scores.

As previously mentioned, the computation of a reference representation in P-ESTOI (independently of whether natural or synthetic speech is used) requires selecting a random initial intelligible representation $\mathbf{H}_r^n$ (cf. Section 6.2.2) from the given set of natural or synthetic utterances. To analyze the sensitivity of P-ESTOI measure to the initial reference representation, we repeat the computation of P-ESTOI multiple times using a different selection of the initial representation for creating the reference representation. The presented correlation values for P-ESTOI measure are the mean and standard deviation of the correlation values obtained for these different repetitions.

To analyze the proposed measures and demonstrate their applicability, the following two scenarios are considered.

Phonetically-balanced scenario

In this scenario, we assume that all speakers (healthy and pathological) utter exactly the same utterances, and the final intelligibility score is computed as the mean across all utterance-level intelligibility scores.

The performance of P-ESTOI in this phonetically-balanced scenario is compared to many blind state-of-the-art measures from Falk et al. (2012), and two non-blind ASR-based measure and the iVector-based measure (Martínez et al., 2015). For the ASR-based and iVector-based approaches, we report the results from Martínez et al. (2015), where these approaches are only evaluated on the English database of CP patients using a leave-one-out validation strategy. For

blind state-of-the-art feature-based methods, we consider several measures which have been shown to yield a high correlation with subjective intelligibility scores in Falk et al. (2012), i.e., the kurtosis of the linear prediction residual $\mathscr{K}_{LP}$, the standard deviation of the zeroth-order delta coefficient σ_{Δ}, the voicing percentage $\%\mathscr{V}$, the range of the fundamental frequency Δ_{f0}, and the low-to-high modulation energy ratio (LHMR). $\mathscr{K}_{LP}$ aims at characterizing vocal source excitation atypicality, σ_{Δ} aims at characterizing short-term temporal dynamics, $\%\mathscr{V}$ and Δ_{f0} aim at characterizing disordered prosody, and LHMR aims at characterizing long-term temporal dynamics. Δ_{f0} and the voicing percentage have been computed using Praat (Boersma, 2002), the linear prediction residual and σ_{Δ} have been computed using the speech signal processing (SSP) Python package (Garner, 2013), and LHMR has been computed using Falk et al. (2010). It should be noted that the used voice activity detector and all implementation details for the different measures have not been reported in Falk et al. (2012), hence, the obtained results evaluated on the same English database are different from the ones reported in Falk et al. (2012).

To provide empirical evidence on the robustness of P-ESTOI measure with healthy speech references to gender and age variations, recordings of healthy speakers from the Spanish PC-GITA database (Orozco-Arroyave et al., 2014a) (cf. Section 2.2.1) are used. We consider recordings of 50 healthy Spanish-speaking speakers (25 males and 25 females) from this database with each speaker uttering 10 sentences. The age of the speakers ranges from 31 to 86 years old, with a median age of 62 years old (Orozco-Arroyave et al., 2014a).

Only in such phonetically-balanced scenarios can the performance of P-ESTOI measure using synthetic speech references be compared to the performance of P-ESTOI using healthy speech references (since otherwise healthy speech reference models cannot be generated). Since only high-quality TTS systems were available for the English language, P-ESTOI with synthetic speech references is only evaluated for the English database. The effect of the number of TTS systems used to generate reference representations is also analyzed in this scenario. When comparing P-ESTOI with natural healthy speech references to its counterpart P-ESTOI with synthetic speech references, we compute reference representations from 4 healthy speakers from the English database, i.e., $R = 4$ (cf. Section 6.2.2), since 4 speaker-dependent TTS systems are trained for creating synthetic speech references. However, in general, for comparing P-ESTOI with natural healthy speech references to other state-of-the-art measures, all available healthy speakers in databases are used to create the intelligible speech references.

Phonetically-unbalanced scenario

In this scenario, we assume that all speakers (healthy and pathological) utter different utterances. P-ESTOI using healthy speech references cannot be used in such scenarios since healthy reference models cannot be generated. Instead, only the performance of P-ESTOI using synthetic speech references is evaluated, since it is applicable to such phonetically-unbalanced scenarios. The effect of the number of utterances that is available to estimate intelligibility is also investigated.

Table 6.1 – Performance of the phonetically-balanced intelligibility assessment on the English CP and Dutch HI databases using the proposed (i.e., P-ESTOI with natural speech references) and state-of-the-art measures. The entry denoted by $\{\cdot\}^*$ indicates non-significant correlation, and entries denoted by {-} indicate that correlation values are not available.

	15 English CP patients		16 Dutch HI patients	
Intelligibility Measures	R	R_S	R	R_S
P-ESTOI$_H$	$\mathbf{0.95} \pm 0.00$	$\mathbf{0.94} \pm 0.00$	$\mathbf{0.80} \pm 0.01$	$\mathbf{0.80} \pm 0.01$
iVector	0.74	-	-	-
ASR	0.55	-	-	-
Δ_{f0}	-0.63	-0.61	-0.08^*	-0.09^*
LHMR	-0.42^*	-0.44^*	-0.36^*	-0.40^*
$\%\mathcal{V}$	-0.37^*	-0.58	-0.33^*	-0.33^*
$\mathcal{K}_{LP}$	0.46	0.49	-0.34^*	-0.16^*
σ_Δ	0.49	0.58	0.08^*	0.04^*

6.3.2 Results

Performance in phonetically-balanced scenario

In this section, the performance of the proposed P-ESTOI measure in phonetically-balanced scenarios is compared to the performance of state-the-art measures. In addition, the robustness of P-ESTOI to gender and age variations is empirically analyzed. The performance of P-ESTOI measure using synthetic speech references as opposed to natural speech references is also investigated.

P-ESTOI and state-of-the-art measures

Table 6.1 presents the Pearson and Spearman correlation values obtained for the CP and HI patients using the proposed and state-of-the-art measures. P-ESTOI intelligibility measure with healthy references is denoted as P-ESTOI$_H$. The Pearson correlation values obtained for the CP patients using the iVector- and ASR-based approaches in Martínez et al. (2015) are also presented. As previously mentioned, only the Pearson correlation coefficients for the CP patients have been reported in Martínez et al. (2015). Hence, results for HI patients and Spearman correlation values for CP patients are not available.

To assess the statistical significance of the reported correlation values, entries in Table 6.1 are compared to the corresponding critical correlation values in Table 3.1 (cf. Section 3.3).

It can be observed that P-ESTOI$_H$ gives the highest (significant) correlation values on both databases (i.e., for both considered languages and diseases). Comparing P-ESTOI$_H$ to other non-blind state-of-the-art iVector- and ASR-based approaches, it can be observed that P-ESTOI$_H$ significantly outperforms them for the CP patients. Considering the rest of state-of-the-art measures, i.e., blind approaches, P-ESTOI$_H$ significantly outperforms them as well.

For the CP patients, some of the blind state-of-the-art measures, e.g., Δ_{f0}, $\mathscr{K}_{LP}$, and σ_{Δ} also yield significant correlations with the subjective intelligibility scores while they do not show significant correlations for Dutch HI patients. The fundamental advantage of P-ESTOI over the blind measures is that it relies on comparing a perceptual representation of the pathological speech to a reference perceptual representation of intelligible (healthy) speech, resulting in a high performance independently of the language or of the disease.

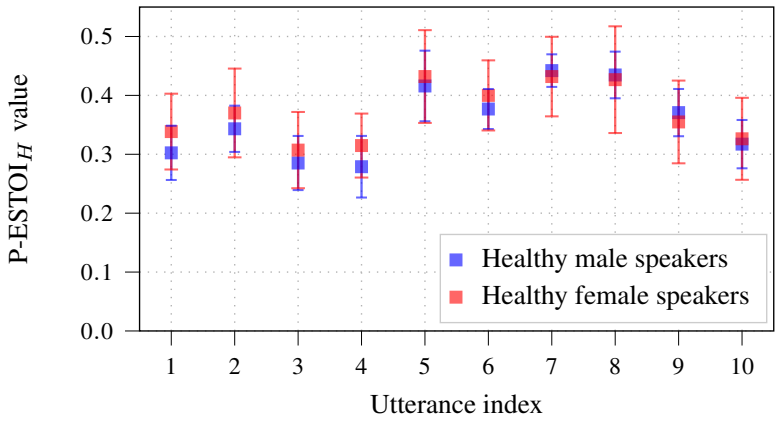
Furthermore, low standard deviations of correlation values obtained by P-ESTOI$_H$ suggest that the computed P-ESTOI$_H$ measure is not sensitive to the selected initial reference representation. In addition, our further experimental results (not presented here) suggest that it is beneficial to use gender-specific reference representations, i.e., reference representation constructed using only healthy male (female) speakers when evaluating the intelligibility of male (female) patients. However, if the number of available healthy speakers is too small, it is more beneficial to use all speakers and create a single reference representation for both male and female patients

Robustness of P-ESTOI to gender and age variations

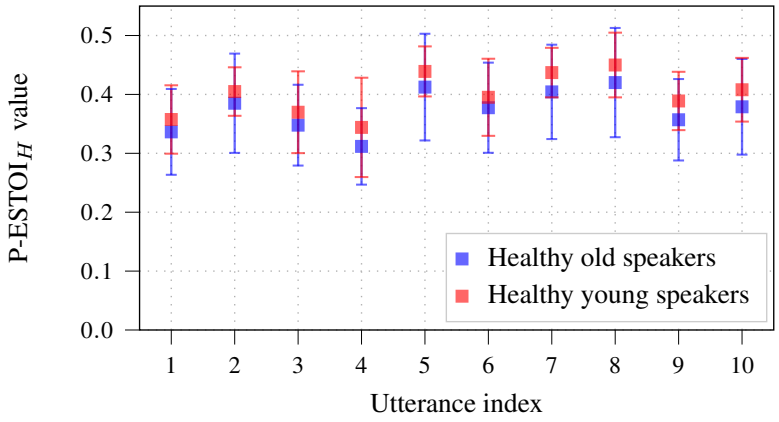
A robust objective intelligibility measure should not be significantly impacted by non-pathological characteristics of speech such as gender- and age-related features. In this section, we investigate the robustness of the proposed P-ESTOI$_H$ measure to the gender and age of speakers. To ensure that the only source of variability is the gender or the age instead of pathology-related features, the following analyses are conducted on healthy (i.e., perfectly intelligible) speech recordings from the PC-GITA database (cf. Section 2.2.1).

To investigate the effect of gender on P-ESTOI$_H$, utterances of 20 (10 males and 10 females) speakers are used to represent the intelligible reference speech signals. To represent the test speech signals, utterances of 30 (15 males and 15 females) speakers are used. The disjoint subsets of intelligible and test speakers are randomly chosen from all available healthy speakers in the PC-GITA database, and the selection of these subsets is repeated 100 times. The P-ESTOI$_H$ measure is then computed for each test utterance from each of the test male and female speakers. For each test utterance, the reference representation is computed as in Section 6.2.2 by DTW-based clustering of the representations of the same utterance from the 20 speakers used as the intelligible reference speakers.

To investigate the robustness of P-ESTOI$_H$ to age, a similar analysis is conducted by dividing the speakers into two age groups (i.e, a young group of speakers with age $\leqslant$ 62 years old and an old group of speakers with age $>$ 62 years old). To represent the intelligible speech signals, utterances of 18 (9 old and 9 young) speakers are used. To represent the test speech signals, utterances of 30 (15 old and 15 young) speakers are used. The disjoint subsets of intelligible and test speakers are also randomly chosen from all available healthy speakers in the PC-GITA database, and the selection of these subsets is repeated 100 times. The P-ESTOI$_H$ measure is then computed for each test utterance from each of the test young and old speakers. For each test utterance, the reference representation is computed as in Section 6.2.2 by DTW-based



(a)



(b)

Figure 6.2 – Mean and standard deviation of the obtained P-ESTOI$_H$ values for 10 utterances across a) male and female speakers, and across b) old and young speakers for one repetition of the speakers' subset selection. For the used speakers' subset selection, no statistically significant differences between mean P-ESTOI$_H$ values of male and female speakers and no statistically significant differences between mean P-ESTOI$_H$ values of old and young speakers for any of the utterances are found.

clustering of the representations of the same utterance from the 18 speakers used as the intelligible reference speakers.

Figs. 6.2a and 6.2b depict the mean and standard deviation of the obtained P-ESTOI$_H$ values for each utterance across the male and female speakers and across the young and old speakers. These results are obtained for one disjoint subset of intelligible and test speakers randomly chosen from all available healthy speakers in the PC-GITA database. It can be observed that the obtained mean P-ESTOI$_H$ values are similar across the two gender and age groups, independently of the considered utterance. This shows that the proposed P-ESTOI$_H$ measure is barely affected by the gender or age of speakers.

To evaluate whether there are significant differences between the mean P-ESTOI$_H$ values for each utterance across the groups of speakers (i.e., male vs. female groups and old vs. young groups), an independent samples t-test is conducted. The t-test is conducted for each repetition of the speakers' subset selection in both gender- and age-based analyses. Out of the 10 considered utterances, the average number of utterances across all repetitions which yields a statistically significant difference (i.e., $p < 0.01$) between the mean P-ESTOI$_H$ values of male and female speakers is less than 1. Similarly, the average number of utterances across all repetitions which yields a statistically significant difference (i.e., $p < 0.01$) between the mean P-ESTOI$_H$ values of old and young speakers is also less than 1. For the speakers' subset selection used in Fig. 6.2, no statistically significant differences for any of the utterances are found. Hence, it can be said that the difference in the mean P-ESTOI$_H$ values across male and female speakers and the difference in the mean P-ESTOI$_H$ values across old and young speakers is generally not statistically significant.

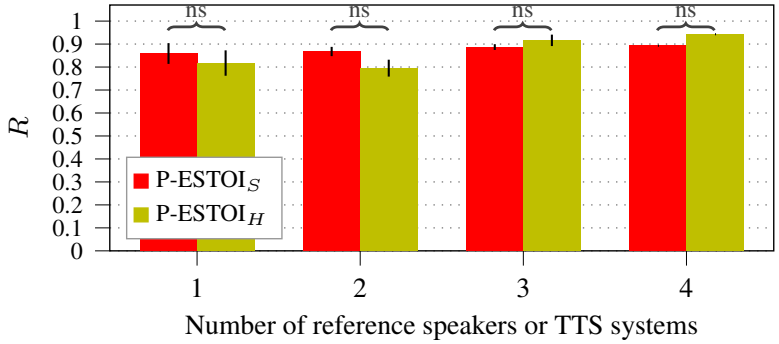
In summary, our analyses show that the proposed P-ESTOI$_H$ measure is not sensitive to the gender and age of speakers and is able to construct representations that reflect intelligibility-related degradations.

Comparing P-ESTOI using synthetic and natural references

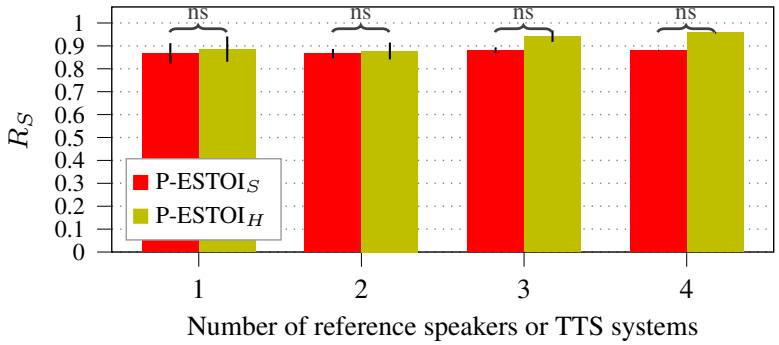
In the following, the performance of P-ESTOI using synthetic speech references is compared to using healthy speech references for assessing the intelligibility of English CP patients. The P-ESTOI intelligibility measure with synthetic references is denoted as P-ESTOI$_S$ and as mentioned before, the P-ESTOI intelligibility measure with natural healthy speech references is denoted as P-ESTOI$_H$. As previously mentioned, P-ESTOI$_S$ is only evaluated for the database with English CP patients, as we only have access to high-quality synthetic English speech samples generated by our trained TTS systems.

To analyze whether the performance of P-ESTOI measure is dependent on the number of TTS systems used to generate the reference representation, we investigate the performance when using 1, 2, 3, and 4 such TTS systems. Accordingly, this is compared to the performance of P-ESTOI using natural speech references generated from 1, 2, 3, and 4 healthy speakers. Since there are multiple ways of selecting, 1, 2, or 3 TTS systems or healthy speakers out of the available 4 TTS systems or healthy speakers, we have repeated the computation of P-ESTOI for each of these possible selections.

Fig. 6.3 presents the Pearson and Spearman rank correlation between the subjective intelligibility scores and the P-ESTOI intelligibility measure using synthetic references (i.e., P-ESTOI$_S$) and healthy references (i.e., P-ESTOI$_H$) for different numbers of TTS systems or healthy speakers. The columns and bars in Fig. 6.3 present the mean and standard deviation of the correlation values across all repetitions. Although not presented in this figure, the correlation values obtained using both P-ESTOI$_S$ and P-ESTOI$_H$ across all repetitions are statistically significant (using the comparison to the corresponding critical correlation values in Table 3.1). It can be observed that the Pearson and Spearman correlation values obtained using P-ESTOI$_S$



(a)



(b)

Figure 6.3 – a) Pearson correlation R and b) Spearman rank correlation R_S using P-ESTOI$_S$ and P-ESTOI$_H$ for different number of TTS systems and healthy speakers. The columns and bars depict the mean and standard deviation of the correlation values across different selections of the set of TTS systems or healthy speakers. (ns) denotes non-significant differences between the correlation values of P-ESTOI$_S$ and P-ESTOI$_H$.

and P-ESTOI$_H$ are both high and very similar, independently of the number of TTS systems or healthy speakers used to generate the reference representations. When using 1 or 2 TTS systems or healthy speakers, the Pearson correlation obtained with P-ESTOI$_S$ is slightly higher than the Pearson correlation obtained with P-ESTOI$_H$. In the remainder of the considered scenarios, the correlation values obtained with P-ESTOI$_S$ are slightly lower than the correlation values obtained with P-ESTOI$_H$. Given all 4 TTS systems used to generate the reference representations, P-ESTOI$_S$ yields mean and standard deviation of R and R$_S$ of 0.89 ± 0.01 and 0.88 ± 0.02 for the English CP patients.

To analyze whether the differences in the presented correlation values of P-ESTOI$_S$ and P-ESTOI$_H$ for each considered number of TTS systems or healthy speakers are statistically significant, we conduct a two-tailed dependent Steiger's Z-Test on all possible pairs of correlation values obtained across all repetitions (Steiger, 1980). The difference between the correlation values is considered significant when the obtained p-value is $p < 0.01$ in the majority (i.e., more than 50%) of the considered correlation pairs, otherwise this difference is

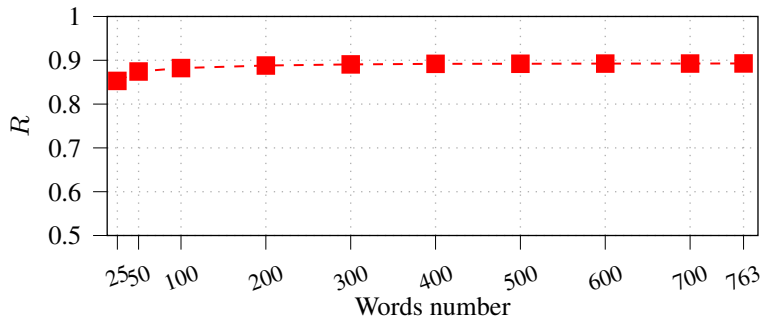
considered to be non-significant (depicted by ns in Fig. 6.3). As shown in Fig. 6.3, there is no significant difference between P-ESTOI$_S$ and P-ESTOI$_H$ independently of the number of TTS systems or healthy speakers.

In summary, it can be said that the performance of P-ESTOI using synthetic speech references is high and similar to using natural speech references.

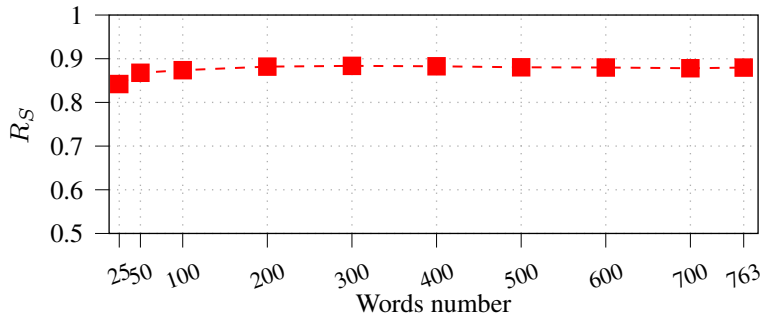
Performance in phonetically-unbalanced scenario

In this section, the performance of P-ESTOI using synthetic speech references is evaluated in phonetically-unbalanced scenarios for the English database (i.e., assessing the speech intelligibility of CP patients).

To investigate the effect of the number of available utterances in estimating intelligibility, a set of utterances is randomly selected from the 763 available utterances for each speaker. The number of considered utterances ranges from 25 to 763. Clearly, there might be common utterances among speakers, however, these utterances are not exactly the same. The random selection of the set of utterances is repeated 100 times, and the final intelligibility score is obtained by averaging across all repetitions.



(a)



(b)

Figure 6.4 – a) Pearson correlation R and b) Spearman rank correlation R_S using P-ESTOI$_S$ in phonetically-unbalanced scenarios for different number of considered utterances.

Fig. 6.4 presents the Pearson and Spearman rank correlation obtained using P-ESTOI$_S$ in the phonetically-unbalanced scenario for different numbers of utterances. Although not presented in this figure, the correlation values obtained using P-ESTOI$_S$ are always statistically significant for each considered number of utterances. It can be observed that independently of the considered number of utterances for intelligibility assessment, the correlation values obtained using P-ESTOI$_S$ are always high, demonstrating the advantages of the proposed method in phonetically-unbalanced scenarios. Further, it can be observed that the correlation values obtained using P-ESTOI$_S$ quickly converge, showing that a relatively small number of utterances is necessary for P-ESTOI$_S$ to obtain a robust intelligibility assessment.

6.4 Summary

In this chapter, we have proposed to automatically assess the intelligibility of pathological speech based on a short-time objective intelligibility measure typically used in speech enhancement, which however requires a reference signal that is time-aligned to the test signal. We have proposed to create an utterance-dependent reference signal of intelligible speech from multiple healthy speakers. In order to assess intelligibility, the pathological speech signal is aligned to the created reference signal using DTW. Our proposed intelligibility measure, i.e., P-ESTOI, is finally computed by quantifying the divergence between the two signals using the spectral correlation. By relying on a reference representation created from multiple healthy speakers, P-ESTOI has shown a high performance (i.e., obtaining high correlation values with subjective intelligibility ratings) independently of the language, i.e., English or Dutch, or the pathology, i.e., CP and HI, and also outperforms the state-of-the-art pathological speech intelligibility measures. In addition, we have shown that the proposed measure is robust to gender- and age-induced changes in the acoustical properties of signals. Hence, our proposed measure is reliable, simple, and does not need a large amount of training data. In addition, it is based on developing a single feature correlated with subjective intelligibility ratings, therefore, it is advantageous over methods that require many features (since training, and hence, overfitting is avoided).

However, to assess the intelligibility of an utterance from a patient, P-ESTOI relies on the availability of recordings of the same utterance by several healthy speakers such that an intelligible reference model can be created. Such recordings are not always easily available, limiting the practical applicability of P-ESTOI. To be able to use P-ESTOI in such scenarios, we have also proposed to use synthetic speech generated by state-of-the-art high-quality TTS systems to create an intelligible reference model. Experimental results on a database of CP patients have shown that the performance of P-ESTOI using synthetic speech references is comparable to using natural speech references. Therefore using synthetic speech references can make P-ESTOI a flexible measure that does not require healthy speech recordings and can be successfully used in a wide range of scenarios while outperforming state-of-the-art pathological speech intelligibility measures.

7 Pathological speech intelligibility assessment exploiting subspace-based analyses

In this chapter we propose a measures to assess the intelligibility of pathological speech based on analyzing the subspaces spanning the dominant speech spectral patterns. This measure unlike our proposed intelligibility measure in the previous chapter does not require any time-alignment, is not computationally expensive, and can also be used in phonetically-unbalanced scenarios. The applicability of the proposed subspace-based intelligibility measure in this chapter is experimentally investigated across databases and compared to many state-of-the-art pathological speech intelligibility measures.

7.1 Introduction

In Chapter 6, we proposed a non-blind pathological speech intelligibility measure based on the extended short-time objective intelligibility, i.e., P-ESTOI measure to tackle the drawbacks of state-of-the-art techniques. P-ESTOI does not rely on extracting a large number of features, does not require any training or a large number of healthy speech recordings, and was shown to be highly correlated with subjective intelligibility ratings for patients suffering from different pathologies. However, for assessing the intelligibility of a sample utterance from a patient in P-ESTOI, recordings of the same utterance from several healthy speakers are needed such that an utterance-dependent reference model can be created. Intelligibility is then assessed through time-alignment of the pathological utterance with the utterance-dependent reference model. Consequently, P-ESTOI measure cannot be used in scenarios where such healthy recordings perfectly matching the phonetic content of the pathological speech signal are not available. To be able to use P-ESTOI measure in such scenarios, in Chapter 6, we also proposed to use synthetic speech generated by state-of-the-art high-quality TTS systems to create an intelligible reference model. However, training TTS systems requires a large amount of healthy speech data with the same language as patients under evaluation, which might not be easily available for under-resourced languages. In addition, the computational cost of time-alignment performed by DTW in P-ESTOI measure (using either natural or synthetic speech references) is high when aligning long utterances.

Aiming to develop an automatic intelligibility measure that does not require any time-alignment and can be used in phonetically-unbalanced scenarios without being computationally expensive, in this chapter we propose a subspace-based intelligibility (SBI) measure. This measure is inspired by the knowledge that speech pathologies typically decrease the degree of spectral modulation in pathological speech signals (Rosen et al., 2006). We hypothesize that pathology-induced spectral modulation changes are reflected in the subspace spanned by the most dominant speech spectral basis vectors. In addition, we hypothesize that the divergence between intelligible and pathological speech spectral subspaces (computed from healthy and pathological speech recordings that do not necessarily share the same phonetic content) can be used as an automatic intelligibility measure.

In this chapter first, we propose to characterize spectral subspaces using a (possibly) different number of spectral basis vectors for the healthy and pathological speech signals. Second, we provide empirical evidence on i) the relation between the SBI measure and low-frequency components of the spectral modulation of speech, which have been shown to be crucial for speech intelligibility, ii) the robustness of the SBI measure to gender variations, and iii) the robustness of the SBI measure to age variations. Third, we provide insights on the computational complexity reduction that is achieved using SBI instead of P-ESTOI measure. Fourth, we propose two techniques to incorporate short-time temporal information in the SBI measure. Finally, we provide an extensive experimental evaluation of the proposed measures to investigate their applicability in phonetically-balanced and phonetically-unbalanced scenarios and their generalisability across languages, i.e., English and Dutch, and across pathologies, i.e., CP and HI. Experimental results show that the proposed measures yield high and significant correlations with subjective intelligibility scores, while not requiring any training or a large number of healthy speech recordings and being applicable to phonetically-unbalanced scenarios.

It should be noted that in Chapter 4 we also used a similar subspace-based characterization of speech signals for automatic pathological speech detection. We extracted both spectral and temporal subspaces spanning the dominant spectro-temporal patterns of speech and used them as acoustic features representing each speaker for the detection task. However, in this chapter, we characterize only the dominant speech spectral basis vectors reflecting pathology-induced spectral modulation changes that are important to the perceived speech intelligibility, and the subspace analysis used for pathological speech intelligibility assessment here is different than subspace-based learning used for the detection task in Chapter 4.

This chapter is organized as follows. Section 7.2 presents a brief overview of the psychoacoustic evidence on the importance of spectral modulation frequencies on speech intelligibility. Section 7.3 describes the proposed SBI measure, and Section 7.4 describes the proposed temporal extensions. Section 7.5 presents empirical insights on the relation between the proposed SBI measure and spectral modulation frequencies and on the robustness of SBI to gender and age variations. Experimental results using the proposed SBI measure and its temporal extensions are presented in Section 7.6, and finally, Section 7.7 presents a summary

of this chapter.

7.2 Modulation spectrum and speech intelligibility

In this section, we first present a brief overview of the psychoacoustic evidence supporting the relation between spectral modulation frequencies and speech intelligibility.

Fluctuations of the speech power spectrogram in time (at any frequency) and in frequency (at any time frame) are referred to as temporal and spectral modulations. Psychoacoustic studies have shown that the temporal and spectral modulations of speech are critical to speech perception, since they represent phonological information such as syllable boundaries and formant information (Drullman et al., 1994; Shannon et al., 1995; Zeng et al., 2005; Elliott and Theunissen, 2009; Hermansky, 2011). The importance of spectro-temporal modulations to speech intelligibility is further confirmed by the success of several objective intelligibility measures typically used in speech enhancement which aim to incorporate (or indirectly assess) modulation cues, such as the speech transmission index (Steeneken and Houtgast, 1980), the spectro-temporal modulation index (Elhilali et al., 2003), LHMR (Falk et al., 2012), envelope power spectrum-based measures (Jorgensen et al., 2013; Biberger and Ewert, 2017), and ESTOI measure (Jensen and Taal, 2016). Since our previously proposed pathological intelligibility measure, i.e., P-ESTOI, is based on ESTOI, it can be easily deduced that P-ESTOI also reflects differences in the spectro-temporal modulation of intelligible and pathological speech. While temporal modulations are indeed very important to speech intelligibility (Drullman et al., 1994; Shannon et al., 1995; Zeng et al., 2005; Elliott and Theunissen, 2009), the objective in this paper is to develop a measure that does not require time-alignment and which can be used in phonetically-unbalanced scenarios. Hence, the proposed SBI measure can only reflect spectral modulation differences between healthy and pathological speech.

The effect of spectral modulation cues on the perceived speech intelligibility by human listeners (i.e., subjective speech intelligibility) has been extensively analyzed in Elliott and Theunissen (2009). In Elliott and Theunissen (2009), the spectral modulation pattern is obtained by computing the Fourier transform of each time frame of the TF representation of utterances. TF representations with a linear frequency axis result in spectral modulations in units of cycle/kHz, whereas TF representations with one-third octave band frequency axis result in spectral modulation in units of cycle/$\frac{1}{3}$octave. To investigate the spectral modulation frequencies contributing to speech intelligibility, the spectral modulation spectrum at each time frame is low-pass filtered at different cut-off frequencies. Using such low-pass filtering, the oscillations in the spectral modulation domain with frequencies above the considered cut-off frequency (i.e., higher-frequency components of the spectral modulation) are removed, while oscillations below the considered cut-off frequency (i.e., lower-frequency components of the spectral modulation) are preserved. The time-domain signal corresponding to the low-pass filtered signal in the spectral modulation domain is reconstructed, and human listeners are asked to rate the intelligibility of these synthetically manipulated utterances.

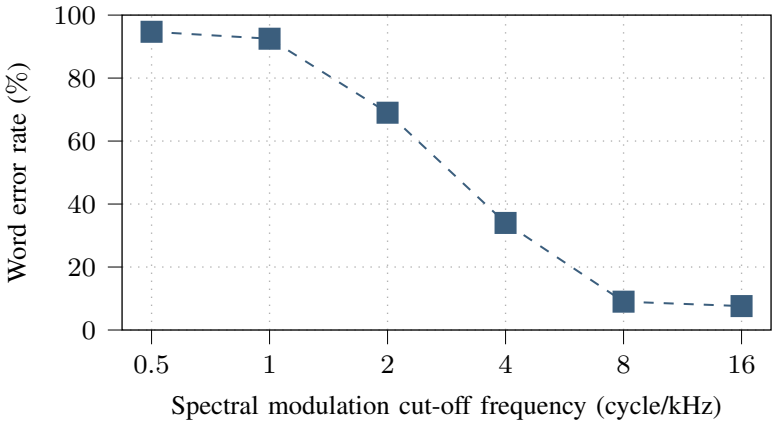


Figure 7.1 – Subjective intelligibility of low-pass spectral modulation filtered utterances based on the percentage of words misunderstood by human listeners. The spectral modulation spectrum of utterances is low-pass filtered at different cut-off frequencies (figure adapted from Elliott and Theunissen (2009)).

Fig. 7.1 shows the effect of low-pass modulation filtering at different cut-off frequencies on the word error rate, i.e., the percentage of words misunderstood by listeners. As it can be observed, the word error rate significantly increases, i.e., speech intelligibility significantly decreases, when low spectral modulation frequencies are missing from the speech signal (Elliott and Theunissen, 2009). Low spectral modulation frequencies represent spectral amplitude fluctuations imposed by the vocal tract, i.e., formants and formant transitions (Elliott and Theunissen, 2009). Hence, it is to be expected that the removal of low spectral modulation frequencies yields a decrease in speech intelligibility. As will be shown in Section 7.5.1, the SBI measure proposed in this paper responds to missing spectral modulation frequencies similarly to Fig. 7.1, i.e., similarly to how humans rate the perceived speech intelligibility when spectral modulation frequencies are missing in the speech signal.

7.3 Subspace-based pathological speech intelligibility assessment

It is commonly accepted that speech spectrograms can be well approximated by low-rank matrices constructed using low-dimensional spectral patterns. Because of the reduced extent of articulatory movements in pathological speakers, the spectral variations in pathological speech are reduced (Rosen et al., 2006). Therefore, it can be expected that the dominant spectral patterns characterizing intelligible (healthy) speech differ from the ones characterizing pathological speech. Hence, we propose to estimate speech intelligibility by quantifying the distance between the spectral subspaces spanned by the dominant spectral basis of pathological speech and the dominant spectral basis of healthy speech. A schematic representation of the proposed SBI measure is depicted in Fig. 7.2. As depicted in this figure, SBI relies on i) computing spectral basis vectors characterizing spectral patterns in intelligible (i.e., healthy) utterances (referred to as intelligible spectral basis vectors), ii) computing spectral

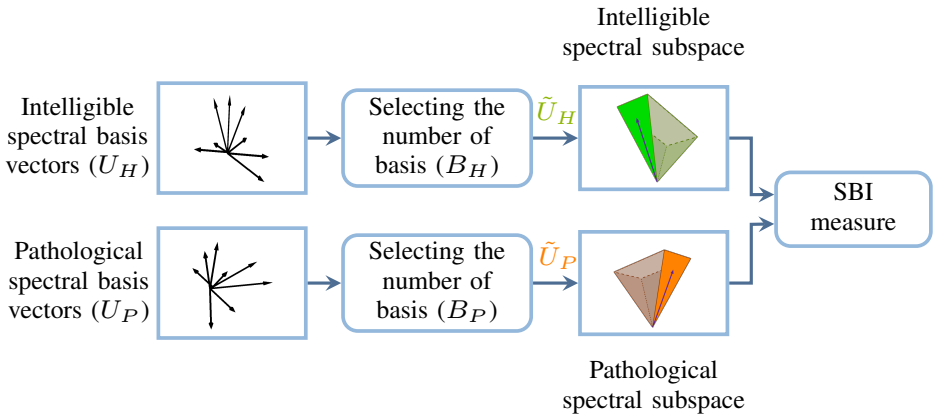


Figure 7.2 – Schematic representation of the proposed subspace-based intelligibility measure. Intelligible and pathological spectral basis vectors are obtained from intelligible (i.e., healthy) and pathological utterances. Low-dimensional spectral subspaces spanned by the most dominant intelligible and pathological spectral basis are created, where the number of dominant spectral basis vectors are automatically found. The pathological intelligibility score is computed as the distance between intelligible and pathological spectral subspaces.

basis vectors characterizing spectral patterns in the test (i.e., pathological) utterance (referred to as pathological spectral basis vectors), iii) automatic selection of the number of spectral basis vectors used to create low-dimensional spectral subspaces corresponding to intelligible and pathological spectral patterns, and iv) computing the intelligibility score as the distance between the intelligible and pathological spectral subspaces. In the remainder of this section, the computational details of the proposed SBI measure are presented.

7.3.1 Computing intelligible spectral basis

While several techniques can be used to compute spectral basis such as approximate joint diagonalization (AJD) (Cardoso and Souloumiac, 1996), non-negative matrix factorization (Lee and Seung, 1999), and sparse coding (Olshausen and Field, 1996), in this paper we propose to use the simple low-rank matrix decomposition minimizing the approximation error in the least-squares sense, i.e., the SVD. The SVD provides an analytical solution and results in a high performance for our application. To obtain meaningful spectral basis vectors, multiple utterances by several healthy speakers should be taken into account, such that the spectral basis vectors can capture patterns that are specific to intelligible speech but are independent of the particular speaker.

To obtain a signal representation resembling the transform properties of the auditory system, signals are first transformed to the TF domain by taking the logarithm of the one-third octave band spectrum (Elliott and Theunissen, 2009; Jensen and Taal, 2016) (cf. Section 2.4.2).

Let $\mathbf{H}_s$ denote the $(J \times M_s)$–dimensional TF representation of an utterance from healthy speaker

s, with J being the total number of one-third octave bands and M_s being the total number of time frames. We consider TF representations of (possibly but not necessarily the same) utterances from different healthy speakers by concatenating them into a $(J \times M)$–dimensional matrix $\mathbf{H}$, where $M = \sum_{s=1}^{S} M_s$, i.e.,

$$\mathbf{H} = [\mathbf{H}_1 \, \mathbf{H}_2 \, \ldots \, \mathbf{H}_S], \tag{7.1}$$

with S being the total number of available healthy speakers. The SVD of $\mathbf{H}$ is given by

$$\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^T, \tag{7.2}$$

with $\mathbf{U}$ being the $(J \times J)$–dimensional orthonormal matrix of left singular vectors representing spectral basis vectors, Σ being the $(J \times M)$–dimensional diagonal matrix of singular values σ_i assumed to be sorted in descending order, and $\mathbf{V}$ being the $(M \times M)$–dimensional orthonormal matrix of right singular vectors. The $(J \times B_H)$–dimensional matrix of dominant intelligible spectral basis vectors $\tilde{\mathbf{U}}_H$ is then constructed from the first B_H (with $B_H < J$) spectral basis vectors in $\mathbf{U}$. The selection of the number of intelligible spectral basis vectors B_H is described in Section 7.3.3.

It should be noted that prior to computing the SVD, the matrices $\mathbf{H}_s$, $s = 1, \ldots, S$, in (7.1) are mean-centered in each octave band and scaled by $\frac{1}{\sqrt{M_s}}$ to remove the bias introduced by the number of time frames. This way, using the SVD in (7.2) to compute the spectral basis vectors is equivalent to PCA (Wall et al., 2003). Although mean-centering the representations in the framework of SVD is optional, it has been shown that the non-zero mean vector across time biases the first spectral basis vector to its direction rather than to the direction with maximal variability of spectral information (Cadima and Jolliffe, 2009; Alexandris et al., 2017).

Instead of using the SVD, we have also investigated the applicability of AJD (Cardoso and Souloumiac, 1996) to extract intelligible spectral basis vectors (the results of which are omitted here). The usage of AJD was motivated by the possibility that spectral subspaces from different healthy speakers might differ significantly. In that case, computing spectral basis vectors by concatenating TF representations from all speakers in (7.1) and then applying SVD in (7.2) might yield basis vectors that do not offer a reasonable approximation to the different representations. Hence, we also proposed to compute the healthy spectral basis vectors by means of AJD (Janbakhshi et al., 2019b). As shown in Janbakhshi et al. (2019b), using the SVD-based decomposition appears to be slightly more advantageous than using the AJD-based decomposition for phonetically-balanced scenarios, while in phonetically-unbalanced scenarios, using the AJD decomposition slightly outperforms using the SVD decomposition. However, in this chapter, we use only SVD-based decomposition to extract basis vectors since, unlike AJD, SVD is simple, provides an analytical solution, and it does not yield a significantly different performance than AJD.

7.3.2 Computing test spectral basis vectors

To be able to assess intelligibility, the test (i.e., pathological) spectral basis vectors also need to be computed. Let $\mathbf{P}_r$ denote the $(J \times M_r)$–dimensional TF representation of the test utterance from patient r, with M_r denoting the total number of time frames. Similar to Section 7.3.1, the SVD of $\mathbf{P}_r$ is computed and the $(J \times J)$–dimensional orthonormal matrix $\mathbf{U}_P$ containing all pathological spectral basis vectors is obtained. Extracting only the dominant B_P basis vectors (with $B_P < J$) from $\mathbf{U}_P$, the $(J \times B_P)$–dimensional matrix of test spectral basis vectors $\tilde{\mathbf{U}}_P$ is constructed. The selection of the number of test spectral basis vectors B_P is described in Section 7.3.3. It should be noted that different from our preliminary research on SBI in Janbakhshi et al. (2019b), to be able to obtain a better approximation of the intelligible and test representations, we allow the number of dominant spectral basis vectors for intelligible and test speech to be different, i.e., $B_H \neq B_P$.

7.3.3 Automatic selection of the number of spectral basis vectors

The number of spectral basis vectors B_H and B_P are hyperparameters of the proposed technique which obviously impact its performance. Using a large number of spectral basis vectors yields a better approximation of the considered TF representations. However, such an approximation is likely to capture not only spectral patterns important to speech intelligibility (i.e., the spectral basis vectors corresponding to larger singular values), but also spectral patterns describing extraneous variations such as speaker variability or noise (i.e., the spectral basis vectors corresponding to smaller singular values). The optimal number of spectral basis vectors should be as small as possible while at the same time it should yield a small approximation error to the original TF representation. Due to this inherent trade-off, in the following, we propose to automatically select the number of spectral basis vectors B_H and B_P by adapting the L-curve method from Hansen (1992), which has been successfully used to automatically select optimal regularization parameters in regularized least-squares techniques (Kodrasi et al., 2013).

To automatically select the number of spectral basis vectors, we propose to use a parametric plot of the approximation error of the original TF representation versus the number of spectral basis vectors. This plot typically has an L-shape, with the corner (i.e., point of maximum curvature) representing a good compromise between the minimization of the approximation error and keeping the number of spectral basis vectors as low as possible. It should be noted that B_H and B_P can also be selected based on a user-defined threshold on the approximation error (as is typically done when using PCA for dimensionality reduction). However, using such a technique requires the user to define a threshold, introducing an additional hyperparameter that needs to be tuned.

The rank-B_H approximation of the original healthy representation $\mathbf{H}$ is obtained using the truncated SVD, i.e.,

$$\hat{\mathbf{H}} = \tilde{\mathbf{U}}_H \tilde{\Sigma}_H \tilde{\mathbf{V}}_H^T, \tag{7.3}$$

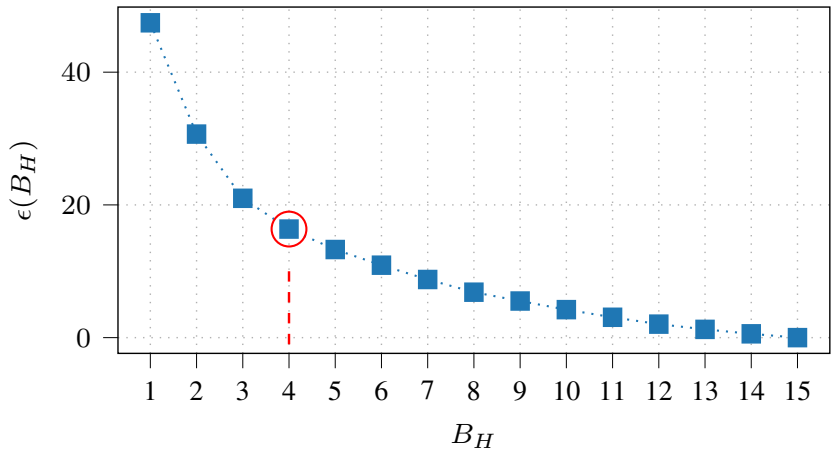


Figure 7.3 – Typical L-curve obtained for the approximation error $\epsilon(B_H)$ versus the number of basis vectors B_H for a sample utterance from the PC-GITA database (cf. Section 2.2.1). The circle depicts the corner point automatically computed using the triangle method.

where $\tilde{\Sigma}_H$ denotes the $(B_H \times B_H)$–dimensional diagonal matrix containing the first B_H singular values and $\tilde{\mathbf{V}}_H$ is the $(B_H \times M)$–dimensional matrix containing the truncated right singular vectors in $\mathbf{V}$. The approximation error $\epsilon(B_H)$ of the intelligible TF representation $\mathbf{H}$ for different number of basis vectors B_H can be computed as

$$\epsilon(B_H) = \left\|\mathbf{H} - \hat{\mathbf{H}}\right\|_F^2 = \sum_{i=B_H+1}^{J} {\sigma_i}^2, \tag{7.4}$$

with $\|\cdot\|_F$ denoting the matrix Frobenious norm and σ_i denoting the i^{th} singular value. The approximation error $\epsilon(B_P)$ of the pathological TF representation $\mathbf{P}_r$ for different number of basis vectors B_P can be computed similarly to (7.4). To automatically select the number of spectral basis vectors B_H and B_P, the parametric plots of $\epsilon(B_H)$ versus B_H and of $\epsilon(B_P)$ versus B_P are constructed. Using the triangle method (Castellanos et al., 2002), the corner points of these parametric plots are computed and used as the number of dominant spectral patterns spanning the intelligible and pathological subspaces.

Fig. 7.3 depicts a typical parametric plot of $\epsilon(B_H)$ versus B_H for a sample utterance from the PC-GITA database (cf. Section 2.2.1). As illustrated in this figure, this parametric plot has an L-shape, with the approximation error $\epsilon(B_H)$ decreasing as the number of spectral basis vectors B_H increases. The corner point automatically computed by the triangle method for this exemplary utterance is also depicted in this figure. Based on the L-curve criterion, using a larger number of basis vectors B_H than the one corresponding to the corner point (i.e., $B_H = 4$ in this example) does not provide any significant reduction in the approximation error. It should be noted that in this work, typical values for the number of basis vectors found with the L-curve method are 3 and 4.

7.3.4 Computing a distance measure between spectral basis vectors

As previously mentioned, the pathological intelligibility score is derived by quantifying the distance between the subspaces spanned by the spectral basis vectors in $\tilde{\mathbf{U}}_H$ (intelligible spectral subspace) and the spectral basis vectors in $\tilde{\mathbf{U}}_P$ (pathological spectral subspace). Since the dimensions of the intelligible and pathological subspaces are typically not the same, i.e., $B_P \neq B_H$, we use a distance measure between subspaces of different dimensions proposed in Ye and Lim (2016). While other subspace distance measures can be used, in this work we use the Procrustes distance defined as

$$\delta(\tilde{\mathbf{U}}_H, \tilde{\mathbf{U}}_P) = 2\sqrt{\sum_{i=1}^{\min(B_H,B_P)} sin^2(\theta_i/2)}, \tag{7.5}$$

where θ_i denotes the i^{th} principal angle between subspaces which can be readily computed via SVD[1] (Ye and Lim, 2016). To be able to compare and combine intelligibility scores from different utterances (i.e., derived from using subspaces of different dimensions), the distance values are normalized to have a maximum value of 1 when the distance between the two subspaces is of the largest possible value, i.e., when $\theta_i = \pi/2,\ i = 1, ..., \min(B_H, B_P)$. Hence, the distance $\delta(\tilde{\mathbf{U}}_H, \tilde{\mathbf{U}}_P)$ obtained for each utterance is scaled by the factor

$$a = \frac{1}{\sqrt{2\min(B_H, B_P)}}. \tag{7.6}$$

It should be noted that the proposed SBI measure is negatively correlated with speech intelligibility since the distance between pathological and intelligible spectral subspaces increases as pathological speech intelligibility decreases. Therefore, such predictions of intelligibility should not be interpreted as an absolute intelligibility score (the percentage of words understood by listeners), but rather should be treated as an index, i.e., expected to be negatively correlated with absolute subjective intelligibility scores. As mentioned in the previous chapter, in this thesis we did not attempt to learn a mapping between predicted intelligibility scores and absolute subjective intelligibility scores.

7.3.5 Complexity analysis

In this section, we provide some insights on the complexity reduction that is achieved when using the SBI measure instead of P-ESTOI proposed in the previous chapter. As mentioned before, computing the spectral basis vectors in (7.2) is equivalent to using PCA on the $J \times J$–dimensional correlation matrix $\mathbf{H}\mathbf{H}^T$. The computation of spectral basis vectors is efficient when PCA is used in practice. Computing correlation matrices requires a complexity of $\mathcal{O}(J^2M)$, where M denotes the number of time-frames (Kwatra and Han, 2010). In addition, the complexity of the PCA decomposition is $\mathcal{O}(J^3)$ (Tammen et al., 2018). Hence, the proposed

[1] It should be noted that the SVD used in Ye and Lim (2016) for computing the principal angles θ_i is unrelated to the SVD in (7.2) representing the spectral basis vectors.

SBI has a computational complexity of $\mathcal{O}(J^2M + J^3)$. The distance computation in (7.5) results in lower-order terms dominated by higher order terms in $\mathcal{O}(J^2M + J^3)$, hence, they are ignored.

When using P-ESTOI, the burden on the computational complexity arises due to using DTW (cf. Section 6.2.2). The DTW algorithm has a computational complexity of $\mathcal{O}(MN)$, with M and N being the number of time frames in the two octave band representations being aligned (Meinard, 2007). Additionally, for each iteration step of DTW, a frame-wise Euclidean distance with complexity $\mathcal{O}(J)$ needs to be computed. Hence, assuming $M = N$, the overall complexity of P-ESTOI is $\mathcal{O}(JM^2)$ (the spectral correlation computation in P-ESTOI (cf. Section 6.2.1) results in lower-order terms dominated by higher order terms in $\mathcal{O}(JM^2)$, hence, they are ignored). Since $M \gg J$ (particularly for long utterances), using the proposed subspace-based measure instead of P-ESTOI reduces the computational complexity by a factor of M (i.e., from $\mathcal{O}(JM^2)$ to $\mathcal{O}(J^2M + J^3)$), which can be advantageous when using such automatic measures for real-time feedback and assistance of clinicians.

7.4 Incorporating temporal information in subspace-based intelligibility measure

The proposed SBI measure in Section 7.3 exploits only the spectral basis vectors in $\mathbf{U}_H$ and $\mathbf{U}_P$ for intelligibility assessment, while ignoring temporal patterns. Although temporal variations are important cues for speech intelligibility, the temporal basis of intelligible and pathological speech cannot be directly computed and compared to each other (because of unaligned and different phonetic contents in the TF representations of intelligible and pathological speech signals). In the following, we propose two viable approaches to incorporate short-time temporal information into the SBI measure. As will be shown in the experimental results in Section 7.6.2, using the proposed approaches to incorporate temporal information in the SBI measure can significantly improve the intelligibility assessment performance.

7.4.1 Dynamic subspace-based intelligibility measure

Motivated by the dynamic PCA approach in Ku et al. (1995); Zhao and Liu (2004), in this section we propose to incorporate short-time temporal information into the SBI measure by modifying the TF representations through concatenating consecutive spectral vectors. Let $\mathbf{h}_m$ denote $(J \times 1)$–dimensional spectral vector at index m of the TF representation $\mathbf{H}$ (i.e., the m^{th} column of $\mathbf{H}$). By concatenating d such consecutive vectors, with d being a user-defined number ($d \ll M$, cf. Section 7.6.1), a new TF representation matrix $\mathbf{H}_{\text{DSBI}}$ is obtained, i.e.,

$$\mathbf{H}_{\text{DSBI}} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_{d+1} & \dots & \mathbf{h}_{(k-1)d+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_d & \mathbf{h}_{2d} & \dots & \mathbf{h}_{kd} \end{bmatrix}, \tag{7.7}$$

where $k = \lfloor \frac{M}{d} \rfloor$. The matrix $\mathbf{H}_{\text{DSBI}}$ in (7.7) is a $(Jd \times k)$–dimensional matrix. The new pathological representation $\mathbf{P}_{\text{DSBI}}$ is obtained similarly to (7.7). Applying the same procedure as for computing the SBI measure in Section 7.3 to the modified representations $\mathbf{H}_{\text{DSBI}}$ and $\mathbf{P}_{\text{DSBI}}$, the dynamic SBI (DSBI) measure of pathological speech intelligibility is obtained. It should be noted that the modified TF representations $\mathbf{H}_{\text{DSBI}}$ and $\mathbf{P}_{\text{DSBI}}$ are of a larger spectral dimension than the original TF representations $\mathbf{H}$ and $\mathbf{P}$ (i.e., the number of rows in $\mathbf{H}_{\text{DSBI}}$ and $\mathbf{P}_{\text{DSBI}}$ is larger than the number of rows in $\mathbf{H}$ and $\mathbf{P}$). Consequently, the number of spectral basis vectors required to span these TF representations is also larger.

7.4.2 Moving average subspace-based intelligibility measure

Motivated by the moving average PCA model in Zhao and Liu (2004), in this section, we propose to incorporate short-time temporal information into the SBI measure by modifying the TF representations through a moving average model. Exploiting a moving average model can account for the short-time temporal correlation of speech signals, which is ignored in the SBI measure. It should be noted that while the DSBI measure proposed in Section 7.4.1 considers multiple time frames simultaneously, the moving average SBI (MASBI) measure proposed in this section considers only a smoothed average across consecutive time frames. Unlike (7.7) where the spectral dimension is increased, the modified TF representation in MASBI has the original spectral dimension of (7.1).

The modified moving average TF representation is constructed as

$$\mathbf{H}_{\text{MASBI}} = [\, \mathbf{h}'_1 \quad \mathbf{h}'_2 \quad \dots \quad \mathbf{h}'_{M-q+1} \,], \tag{7.8}$$

where $\mathbf{h}'_m = \frac{1}{q} \sum_{j=m}^{m+q-1} \mathbf{h}_j$ for $m = 1, \dots, M-q+1$ and q is a user-defined number of time frames (cf. Section 7.6.1). The matrix $\mathbf{H}_{\text{MASBI}}$ in (7.8) is a $(J \times (M-q+1))$–dimensional matrix. The new pathological representation $\mathbf{P}_{\text{MASBI}}$ is also obtained similarly to (7.8). Applying the same procedure as for computing the SBI measure in Section 7.3 to the modified representations $\mathbf{H}_{\text{MASBI}}$ and $\mathbf{P}_{\text{MASBI}}$, the MASBI measure of pathological speech intelligibility is obtained.

7.5 Empirical insights into the proposed subspace-based intelligibility measure

The objective of this section is to show through empirical analyses that the proposed SBI measure focuses on low-frequency spectral modulation cues to assess pathological speech intelligibility. This property can be justified by the psychoacoustic evidence confirming that low-frequency spectral modulations contribute to the perceived speech intelligibility by human listeners (cf. Section 7.2). In addition, we provide empirical evidence on the robustness of SBI to gender and age variations. We used the same protocol as our previously mentioned empirical analysis on the robustness of P-ESTOI to gender and age in Chapter 6

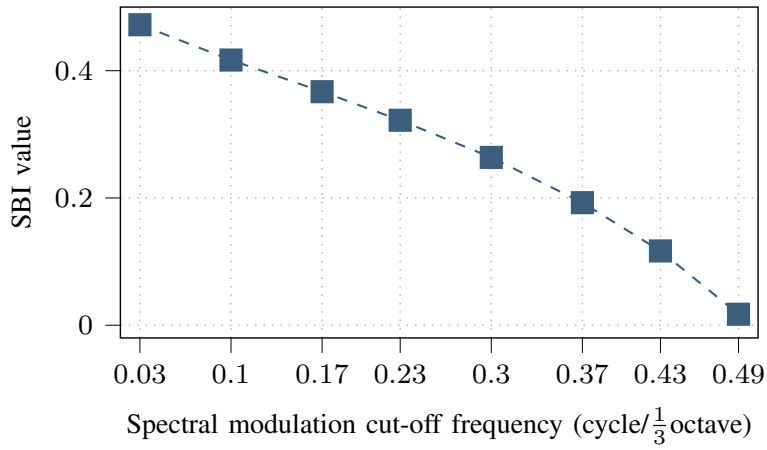


Figure 7.4 – Automatically estimated intelligibility using the proposed SBI measure for low-pass spectral modulation filtered utterances. The cut-off frequency units are cycle/$\frac{1}{3}$octave. The lack of low-frequency spectral modulations of speech has a similar effect on the estimated intelligibility by SBI as on the subjective intelligibility perceived by human listeners as depicted in Fig. 7.1.

(cf. Section 6.3.2). For these analyses, the algorithmic settings described in Section 7.6.1 and recordings of healthy speakers from the PC-GITA database (Orozco-Arroyave et al., 2014a) (cf. Section 2.2.1) are used. Similarly to analyses in Section 6.3.2, we consider recordings of 50 healthy Spanish-speaking speakers (25 males and 25 females) from this database with each speaker uttering 10 sentences. The age of the speakers ranges from 31 to 86 years old, with a median age of 62 years old (Orozco-Arroyave et al., 2014a).

7.5.1 Subspace-based intelligibility measure and spectral modulation of speech

In analogy to the experiment conducted in Elliott and Theunissen (2009) (cf. Section 7.2), in this section, we analyze the effect of spectral modulation cues on the proposed SBI measure. To this end, the modulation spectrum obtained from the TF representation of each utterance from the PC-GITA database is low-pass filtered at different cut-off frequencies. Instead of asking human listeners to evaluate the perceived intelligibility of the low-pass spectral modulation filtered signals as in Elliott and Theunissen (2009), we compute the proposed SBI measure based on the spectral basis vectors spanning the original utterances (representing e.g. healthy speech signals) and the low-pass filtered utterances (representing e.g. pathological speech signals).

Fig. 7.4 depicts the mean intelligibility estimated using the proposed SBI measure across all considered low-pass spectral modulation filtered utterances for different cut-off frequencies.[2]

[2]It should be noted that the cut-off frequencies we use differ from Elliott and Theunissen (2009) due to differences in the parameters of the TF representations. In addition, while Elliott and Theunissen (2009) uses a linear frequency representation resulting in units of cycle/kHz, we use a logarithmic frequency representation resulting in units of cycle/$\frac{1}{3}$octave

It can be observed that the effect of missing spectral modulation frequencies on SBI is similar to Fig. 7.1, i.e., similar to the effect of missing spectral modulation frequencies on the subjective speech intelligibility perceived by human listeners. In other words, the lack of low-frequency modulations in speech signals decreases the intelligibility estimated through the proposed SBI measure in a similar trend to how the perceived intelligibility by human listeners decreases. This observation shows that low-frequency components of spectral modulations are crucial for speech intelligibility assessment through SBI as they are also crucial for the perceived speech intelligibility by human listeners (cf. Section 7.2). This observation is expected since the dominant spectral basis vectors obtained by SVD usually span low-frequency spectral patterns. Consequently, the manipulation of these spectral patterns will be reflected in the proposed SBI measure.

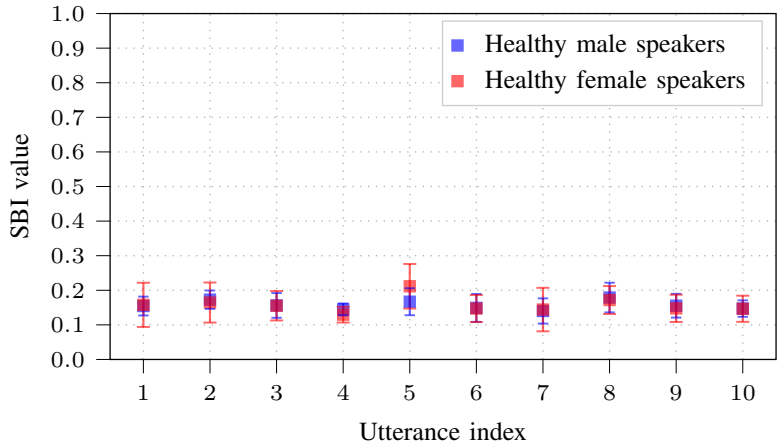
7.5.2 Robustness of the subspace-based intelligibility measure to gender and age variations

To ensure that our proposed SBI measure is not significantly impacted by non-pathological characteristics of speech such as gender- and age-related features, in this section, we investigate its robustness to the gender and age of speakers.

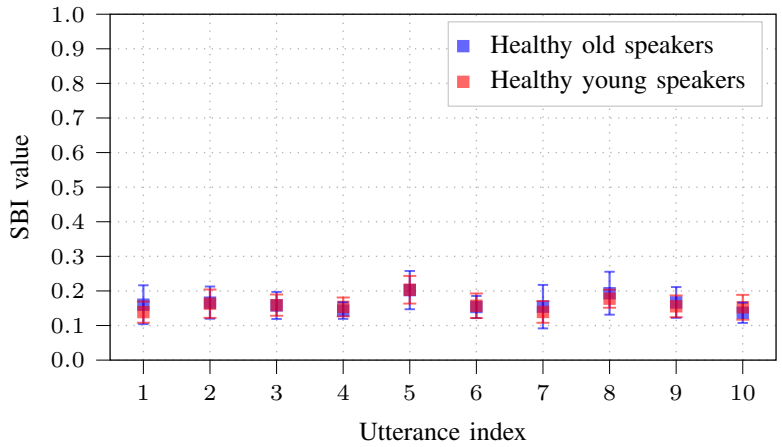
To investigate the effect of gender on SBI, utterances of 20 (10 males and 10 females) speakers are used to represent the intelligible speech signals. To represent the test speech signals, utterances of 30 (15 males and 15 females) speakers are used. The disjoint subsets of intelligible and test speakers are randomly chosen from all available healthy speakers in the PC-GITA database, and the selection of these subsets is repeated 100 times. The SBI measure is then computed for each test utterance from each of the test male and female speakers. For each test utterance, the healthy TF representation is computed as in (7.1) by concatenating multiple instances of this utterance from the 20 speakers representing the intelligible speakers.

To investigate the robustness of SBI to age, a similar analysis is conducted by dividing the speakers into two age groups (i.e., a young group of speakers with age $\leqslant 62$ years old and an old group of speakers with age > 62 years old). To represent the intelligible speech signals, utterances of 18 (9 old and 9 young) speakers are used. To represent the test speech signals, utterances of 30 (15 old and 15 young) speakers are used. The disjoint subsets of intelligible and test speakers are also randomly chosen from all available healthy speakers in the PC-GITA database, and the selection of these subsets is repeated 100 times. The SBI measure is then computed for each test utterance from each of the test young and old speakers. For each test utterance, the healthy TF representation is computed as in (7.1) by concatenating multiple instances of this utterance from the 18 speakers representing the intelligible speakers.

Figs. 7.5a and 7.5b depict the mean and standard deviation of the obtained SBI values for each utterance across the male and female speakers and across the young and old speakers. These results are obtained for one disjoint subset of intelligible and test speakers randomly chosen from all available healthy speakers in the PC-GITA database. It can be observed



(a)



(b)

Figure 7.5 – Mean and standard deviation of the obtained SBI values for 10 utterances across a) male and female speakers, and across b) old and young speakers for one repetition of the speakers' subset selection. For the used speakers' subset selection, no statistically significant differences between mean SBI values of male and female speakers and no statistically significant differences between mean SBI values of old and young speakers for any of the utterances are found.

that the obtained mean SBI values are very similar across the two gender and age groups, independently of the considered utterance. This shows that the proposed SBI measure is barely affected by the gender or age of speakers. In addition, it can be observed that the mean SBI values for all groups of speakers and for all utterances are typically low. This is to be expected since the test signals are perfectly intelligible independently of the gender or age of speakers and the distance between spectral subspaces of intelligible speech should be minimal.

To evaluate whether there are significant differences between the mean SBI values for each

utterance across the groups of speakers (i.e., male vs. female groups and old vs. young groups), an independent samples t-test is conducted. The t-test is conducted for each repetition of the speakers' subset selection in both gender- and age-based analyses. Out of the 10 considered utterances, the average number of utterances across all repetitions which yield a statistically significant difference (i.e., $p < 0.01$) between the mean SBI values of male and female speakers is less than 1. Similarly, the average number of utterances across all repetitions which yield a statistically significant difference (i.e., $p < 0.01$) between the mean SBI values of old and young speakers is also less than 1. For the speakers' subset selection used in Fig. 7.5a and Fig. 7.5b, no statistically significant differences for any of the utterances are found. Hence, it can be said that the difference in the mean SBI values across male and female speakers and the difference in the mean SBI values across old and young speakers is generally not statistically significant.

In summary, our analyses show that the proposed SBI measure is not sensitive to the gender and age of speakers and is able to construct representations that can mainly reflect intelligibility-related degradations.

7.6 Experimental results

In this section, the performance of the proposed intelligibility measures is extensively investigated and compared to state-of-the-art measures. To demonstrate the applicability of the proposed measures for several languages and pathologies, we evaluate the performance on databases of English-speaking CP patients and Dutch-speaking HI patients (cf. Section 3.2). To demonstrate the applicability of the proposed measures for a wide range of scenarios, we consider both phonetically-balanced and phonetically-unbalanced scenarios.

7.6.1 Algorithmic settings, state-of-the-art measures, scenarios, and evaluation

In this section, we present the algorithmic settings for the implementation of the proposed intelligibility measures. The considered scenarios and the performance evaluation metrics are also presented. The considered state-of-the-art measures that are compared to the proposed measures in this chapter are the same as in Chapter 6, i.e., Section 6.3.2.

To compute intelligible representations for the proposed measures, we use speech signals from both healthy male and female speakers. Intelligible representations for the CP patients are constructed using the speech signals of 9 male and 4 female healthy speakers from the English (i.e., UA-Speech database). Intelligible representations for the HI patients are constructed using the speech signals of 11 male and 11 female healthy speakers from the Dutch (i.e., COPAS) database. To obtain the octave-band representations, the same STFT and octave band settings as in the previous chapter are used (cf. Section 6.3.1). The empirically selected number of time frames used to incorporate temporal information in the DSBI and MASBI measures is $d = 5$ (cf. (7.7)) and $q = 9$ (cf. (7.8)), respectively.

As mentioned before, the computation of spectral basis vectors for the proposed measures is efficient when PCA is used in practice. For the SBI and MASBI measures, spectral basis vectors are obtained by applying PCA on the 15×15–dimensional correlation matrices $\mathbf{H}\mathbf{H}^T$ and $\mathbf{H}_{\text{MASBI}}\mathbf{H}_{\text{MASBI}}^T$. Running PCA for such matrices on a computer with a 2.7 GHz processor and 8 GB RAM requires only 0.0003 seconds. For the DSBI measure, spectral basis vectors are obtained by applying PCA on the 75×75–dimensional matrix $\mathbf{H}_{\text{DSBI}}\mathbf{H}_{\text{DSBI}}^T$ (since $d = 5$). Running PCA for such matrices on a computer with a 2.7 GHz processor and 8 GB RAM requires only 0.003 seconds.

The performance of the proposed measures is compared to the performance of the non-blind state-of-the-art measures mentioned in the previous chapter, i.e., iVector-based and ASR-based approaches (Martínez et al., 2015), and our previously proposed P-ESTOI measure (cf. Section 6.2.2). For the iVector- and ASR-based approaches, we report the results from Martínez et al. (2015) where these approaches are evaluated only on the UA-Speech database following a leave-one-subject-out validation strategy.

The performance of the considered intelligibility measures is evaluated for the following two scenarios.

Phonetically-balanced scenarios

In these scenarios, we assume that all speakers (healthy and pathological) utter exactly the same words. All 763 available words are considered for the UA-Speech database, and all 47 available words are considered for the COPAS database. The intelligibility score is calculated for each word, and the final intelligibility score is computed as the mean intelligibility score across all words. Only in such phonetically-balanced scenarios can the performance of the proposed measures be compared to the performance of P-ESTOI (since otherwise healthy speech reference models for P-ESTOI cannot be constructed). In addition, the performance of the iVector- and ASR-based approaches in Martínez et al. (2015) has been analyzed also in such a phonetically-balanced scenario (only for the UA-Speech database).

Phonetically-unbalanced scenarios

In these scenarios, the applicability of the proposed measures is analyzed in the presence of phonetic variability in the considered speech signals from each speaker. Since speakers utter different words in such scenarios, a robust spectral subspace can only be constructed when longer utterances (i.e., longer than a single word) are taken into account. Different sets of words are concatenated to create longer utterances for each speaker, and a single intelligibility score is estimated for each patient. Since the UA-Speech database contains a large number of words that can be combined in different ways for different speakers, these analyses are done on the UA-Speech database. We assess the effect of different levels of phonetic variability on the proposed intelligibility measures by concatenating multiple words for each speaker in the

following manners.

i) The phonetic content within the speakers in each group is the same, while the phonetic content across the two groups of speakers is partially different. To generate this scenario, the set UW is randomly divided into two subsets of equal size (149 words). The utterance uttered by all healthy speakers is created by concatenating one such subset of UW and one repetition (155 words) of the set CW. The utterance uttered by all pathological speakers is created by concatenating the other subset of UW and one repetition (155 words) of the set CW. The total number of concatenated words in each utterance is 304.

ii) The phonetic content within the speakers in each group is the same, while the phonetic content across the two groups of speakers is completely different. To generate this scenario, a similar procedure as in i) is followed. Differently from i), the set CW is also randomly divided into two disjoint subsets (of size 77 and 78 words). The utterance uttered by all healthy speakers is created by concatenating the previously considered subset of UW and one such subset of CW. The utterance uttered by all pathological speakers is created by concatenating the previously considered subset of UW and the other subset of CW. The total number of concatenated words for each healthy speaker is 226, whereas the total number of concatenated words for each pathological speaker is 227.

iii) The phonetic content across all speakers is partially different. To generate this scenario, the utterance for each speaker is created by concatenating 200 randomly selected words from the UW and CW sets. Since there are only a total of 763 words available, there is a partial overlap between the phonetic content across the different speakers.

iv) The phonetic content across all speakers is completely different. To generate this scenario, the utterance for each speaker is created by concatenating 16 distinct (and randomly selected) words from the UW and CW sets.

The subset of words to be concatenated for creating longer utterances for each speaker in the above-mentioned scenarios is randomly selected. This selection is repeated 100 times, and the performance of the proposed measures is analyzed in terms of the mean and standard deviation of the performance across all repetitions.

Similar to the previous chapter, the Pearson correlation coefficient (R) and the Spearman rank correlation coefficient (R_S) between the automatically estimated intelligibility and the subjective intelligibility scores of the CP patients and HI patients are computed.

Table 7.1 – Performance of the phonetically-balanced intelligibility assessment on the English CP and Dutch HI databases using the proposed (i.e., SBI, DSBI, and MASBI) and state-of-the-art (i.e., P-ESTOI, iVector, and ASR) measures. The entry denoted by $\{\cdot\}^*$ indicates non-significant correlation, and entries denoted by {-} indicate that correlation values are not available.

	15 English CP patients		16 Dutch HI patients	
Measures	R	R_S	R	R_S
P-ESTOI	0.95	0.94	0.80	0.80
iVector	0.74	-	-	-
ASR	0.55	-	-	-
SBI	−0.86	−0.88	−0.48	-0.40^*
DSBI	−0.86	−0.93	−0.64	−0.60
MASBI	−0.82	−0.88	−0.68	−0.65

7.6.2 Results

Performance in phonetically-balanced scenarios

In this section, the performance of the proposed measures in phonetically-balanced scenarios is compared to the performance of state-the-art measures.

Table 7.1 presents the Pearson and Spearman correlation values obtained for the CP and HI patients using the proposed measures and the P-ESTOI measure. In addition, the Pearson correlation values obtained for the CP patients using the iVector- and ASR-based approaches in Martínez et al. (2015) are also presented. As previously mentioned, only the Pearson correlation coefficients for the CP patients have been reported in Martínez et al. (2015). Hence, results for HI patients and Spearman correlation values for CP patients are not available. To assess the statistical significance of the reported correlation values, entries in Table 7.1 are compared to the corresponding critical correlation values in Table 3.1 (cf. Section 3.3).

It can be observed that P-ESTOI still gives the highest correlation values on both databases, which is to be expected since P-ESTOI takes both the temporal and spectral distortions into account by aligning the pathological speech signals to the intelligible reference representations. However, this limits the application of P-ESTOI to only such phonetically-balanced scenarios. For the CP patients, the proposed SBI, DSBI, and MASBI measures also yield very high and significant correlations with the subjective intelligibility scores, significantly outperforming the state-of-the-art iVector- and ASR-based approaches. In comparison to the SBI measure, incorporating short-time temporal information as in the DSBI measure slightly increases the obtained correlation on this database. Incorporating short-time temporal information as in the MASBI measure slightly decreases the Pearson correlation coefficient, whereas the Spearman rank correlation coefficient is the same as for the SBI measure. However, the SBI measure does not show significant Spearman rank correlation values on the HI database. Incorporating short-

time temporal information through the DSBI and MASBI measures significantly improves the performance over the SBI measure on this database.

It should be noted that the results presented here are obtained using an arbitrarily selected subset of healthy speakers to generate intelligible representations. We have additionally investigated the sensitivity of the proposed measures to the choice of healthy speakers for computing intelligible representations. Although we have omitted these results here, they show that the performance of the proposed measures is insensitive to the specific healthy speakers used for generating reference representations. Additionally, one can compare the proposed measures to the blind state-of-the-art blind intelligibility measures computed in the previous chapter (cf. Section 6.3.2). Since these measures result in a worse performance than the proposed measures, their results are not repeated here for the sake of brevity.

In summary, it can be said that the proposed measures are applicable to phonetically-balanced scenarios and result in high and significant correlations with subjective intelligibility scores. In addition, it can be said that incorporating short-time temporal information (i.e., as in the DSBI and MASBI measures) can yield a substantial performance improvement as opposed to considering only spectral information (i.e., as in the SBI measure).

Performance in phonetically-unbalanced scenarios

In this section, the performance of the proposed measures is analyzed in phonetically-unbalanced scenarios. It should be noted that the P-ESTOI measure is inapplicable to such scenarios since the phonetic content among all speakers should be the same to be able to create an intelligible reference representation.

Table 7.2 presents the mean and standard deviation of the Pearson and Spearman rank correlation values across all repetitions of words' subset selection obtained using the proposed SBI, DSBI, and MASBI measures for all considered phonetically-unbalanced scenarios. To assess the statistical significance of the reported correlation values, entries in Table 7.2 are compared to the corresponding critical correlation values in Table 3.1 (cf. Section 3.3). Overall it can be observed that all proposed measures typically yield high and significant correlations with the subjective intelligibility scores. In addition, the performance of individual measures for scenarios i)–iii) is very similar, showing that the different levels of phonetic variability in these scenarios do not substantially affect the performance of the proposed measures. However, it can be observed that the performance of the proposed measures for scenario iv) is lower than for the other scenarios. This performance degradation in scenario iv) is to be expected since intelligibility is assessed using only 16 words which are different across all speakers. Such a small number of words with different phonetic content does not suffice to construct a robust subspace reflecting speech intelligibility. While the performance of all proposed measures decreases in this scenario, the performance of the proposed DSBI measure is particularly lower. The DSBI measure relies on a TF representation of a larger spectral dimension than the SBI and MASBI measures. The number of spectral basis vectors required to span the

Table 7.2 – Performance of the phonetically-unbalanced intelligibility assessment on the English CP database using the proposed measures. The entries denoted by $\{\cdot\}^*$ indicate non-significant correlations.

Measures	R	R_S
Phonetically-unbalanced scenario i)		
SBI	-0.74 ± 0.02	-0.76 ± 0.03
DSBI	-0.69 ± 0.05	-0.71 ± 0.06
MASBI	-0.73 ± 0.04	-0.76 ± 0.06
Phonetically-unbalanced scenario ii)		
SBI	-0.73 ± 0.03	-0.75 ± 0.04
DSBI	-0.70 ± 0.06	-0.73 ± 0.06
MASBI	-0.71 ± 0.06	-0.74 ± 0.07
Phonetically-unbalanced scenario iii)		
SBI	-0.73 ± 0.03	-0.76 ± 0.04
DSBI	-0.69 ± 0.06	-0.72 ± 0.07
MASBI	-0.72 ± 0.05	-0.75 ± 0.06
Phonetically-unbalanced scenario iv)		
SBI	-0.70 ± 0.07	-0.72 ± 0.08
DSBI	$-0.37^* \pm 0.16$	$-0.41^* \pm 0.17$
MASBI	-0.65 ± 0.11	-0.65 ± 0.12

intelligible and test representations for this measure is larger. Consequently, to construct robust subspaces when the phonetic content among speakers differ, longer utterances are required for this measure than for the SBI and MASBI measures.

In summary, it can be said that the proposed measures are applicable to phonetically-unbalanced scenarios and result in high and significant correlations with subjective intelligibility scores. Since the phonetic content across speakers differs in such scenarios, incorporating short-time temporal information (i.e., as in the DSBI and MASBI measures) does not yield a performance improvement as opposed to considering only spectral information (i.e., as in the SBI measure).

The presented analyses show the successful applicability of the proposed measures on speech disorders arising due to CP and HI. To the best of our knowledge, a systematic comparison of spectral modulation changes across different pathologies has never been done. If the induced spectral modulation changes are dependent on the pathology, it can be expected that the proposed measures perform differently on different pathologies.

7.7 Summary

In this chapter, we have proposed the automatic pathological speech intelligibility SBI measure, which is based on the assessment of the distance between subspaces spanned by dominant spectral patterns of intelligible (i.e., healthy) and pathological speech. This measure unlike our proposed P-ESTOI intelligibility measure in the previous chapter does not require any time-alignment and can also be used in phonetically-unbalanced scenarios. Exploiting psychoacoustic evidence on the importance of spectral modulation cues to the perceived speech intelligibility, we have shown that the proposed SBI measure is advantageous since it can capture pathology-induced distortions in the spectral modulation cues. In addition, we have shown that the proposed measure is robust to gender- and age-induced changes in the acoustical properties of signals, while also being more computationally efficient than our previously proposed P-ESTOI measure. To be able to additionally track possible degradations in the temporal structure of the pathological speech signal, we have also proposed two extensions of the SBI measure, i.e., the DSBI and MASBI measures. Experimental results for different languages and speech pathologies have shown that the proposed measures obtain high correlations with subjective intelligibility scores, with the incorporation of temporal information into the DSBI and MASBI measures yielding a better performance in phonetically-balanced scenarios. In addition, it has been shown that the proposed measures outperform several non-blind state-of-the-art measures, while not requiring regression training, a large amount of healthy speech training data, time-alignment, and being applicable to phonetically-unbalanced scenarios.

8 Toward a clinical tool for joint automatic speech pathology detection and speech intelligibility assessment

In this chapter we jointly validate pathological speech detection and intelligibility assessment tasks using two of our proposed approaches. The goal of the joint analysis of the two tasks is to confirm the possibility of developing a multi-purpose clinical tool useful for clinicians to perform automatic pathological speech assessment.

8.1 Introduction

As mentioned in Chapter 1, to diagnose speech disorders, speech screening through clinical auditory-perceptual assessments is typically used. For the management and treatment of speech disorders to improve the communication ability of patients, speech intelligibility assessment is also performed in clinical settings. Such clinical approaches to pathological speech detection and intelligibility assessment can be time-consuming and inconsistent, since they are subjective and influenced by the level of expertise of clinicians. Furthermore, subjective intelligibility assessment can be biased by the availability of syntactic/semantic clues in the speech of the speaker under evaluation as well as by the familiarity of the clinician with the speaker or speech disorder. To assist clinical speech screenings, automatic techniques offering objective and repeatable pathological speech assessments with the capability of being used in real-time and also in remote speech therapy applications can be exploited. The high-level objective of this thesis is proposing a multi-purpose automatic tool that can be used by clinicians to evaluate the two aspects of clinical analysis automatically, i.e., pathological speech detection and intelligibility assessment. In this chapter, we aim to jointly validate one of our previously proposed automatic speech pathology detection approaches along with one of our previously proposed intelligibility measures in a unified scenario to be viewed as an evaluation of such a multi-purpose clinical tool.

A schematic representation of the joint analysis for the clinical tool is depicted in Fig. 8.1. As depicted in this figure, such a clinical tool consists of two separate modules for speech pathology detection and intelligibility assessment. The goal is to assist clinicians by automatically

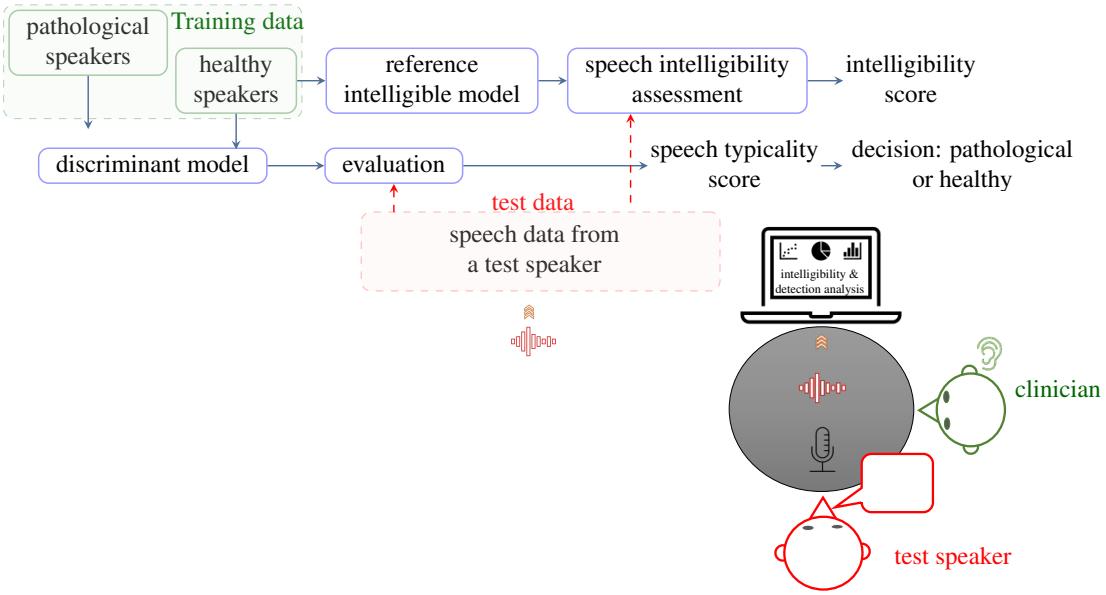


Figure 8.1 – Schematic representation of the clinical tool for joint automatic speech pathology detection and speech intelligibility assessment.

discriminating speakers with atypical speech from speakers with typical speech, followed by further predicting how much the atypical speech intelligibility is affected.

In Chapters 4 and 5, we proposed methods for speech pathology detection, while in Chapters 6 and 7, we proposed suitable intelligibility measures applicable for different scenarios. For the speech intelligibility assessment module of the joint analysis in this chapter, while any of our proposed intelligibility measures can be used, we choose the P-ESTOI intelligibility measure with healthy natural speech reference (cf. Section 6.2.2), as its performance was superior compared to other measures under phonetically-balanced scenarios. It should be noted that due to the need for availability of the ground truth (i.e., subjective intelligibility score) for the speech intelligibility assessment module of the joint analysis in this chapter, we use two databases that were previously considered for intelligibility assessment, i.e., the English UA-Speech database and the Dutch COPAS database (cf. Section 3.2). These databases for intelligibility assessment have a smaller size compared to databases previously used for evaluating the detection task (cf. Section 2.2). Hence, for the detection module of the clinical tool, we choose an approach among our previously proposed ones that has shown a high performance in phonetically-balanced scenarios without requiring a large amount of data, i.e., the temporal subspace-based discriminant approach (i.e., T-GDA) (cf. Chapter 4). It should be noted that to be able to jointly evaluate both modules on these databases, a different evaluation scenario than the one in Chapter 6 needs to be defined (cf. Section 8.2.1). Therefore, the presented P-ESTOI results in the following differ from the ones in Chapter 6. Clearly, the T-GDA results presented in the following also differ from the ones in Chapter 4, since different databases are used.

The remainder of this chapter is organized as follows. Section 8.2 provides experimental results for the pathological speech assessment tool, whereas Section 8.3 presents a summary of this chapter.

8.2 Experimental results

In this section, the performance of the T-GDA approach for pathological speech detection and the performance of the P-ESTOI for intelligibility assessment under a unified scenario is investigated. For the T-GDA approach in this chapter, similarly to the P-ESTOI measure we consider using one-third octave band representations (cf. Chapters 4 and 6).

8.2.1 Evaluation protocol

We consider two databases for the joint analysis, i.e., the English UA-Speech database including recordings from English-speaking healthy speakers and CP patients and the Dutch COPAS database including recordings from the Dutch-speaking healthy speakers and patients with HI (cf. Section 3.2). For both tasks, we consider a phonetically-balanced scenario, where we use the same sets of word utterances uttered by both groups of speakers (healthy and pathological) in each database. As mentioned in Chapter 4, the T-GDA approach requires an initial (arbitrarily selected) healthy reference speaker for the time-alignment. Hence, from the UA-Speech database we consider all 15 CP patients (11 males, 4 females) and 12 healthy speakers (8 males, 4 females) for jointly evaluating the tasks, while one healthy speaker (one male) is used as the initial reference for the time-alignment step. Although many word utterances are available for the UA-Speech database, we randomly select 47 words from the uncommon words (UW) words set for each speaker. Given the small number of speakers in the UA-Speech database, the validation for this database is based on a leave-one speaker-out strategy. From the COPAS database, we consider 16 patients with HI (6 males, 10 females) and 16 (8 males, 8 females) healthy speakers. Furthermore, one more healthy speaker (one male) is used as the initial reference for the time-alignment step. All 47 available words from all speakers are considered from this database. The validation strategy for the COPAS database is a stratified speaker-independent 8-fold cross-validation. As mentioned in Section 6.2.2, P-ESTOI is a reference-based intelligibility measure, i.e., requires data from multiple healthy speakers to create an utterance-dependent reference representation. For creating the reference representations in P-ESTOI measure for each database, we use the healthy speakers in the training set. The speech pathology detection score and the intelligibility score are calculated for each word from the test speakers, and the final detection and intelligibility scores are computed as the mean detection score and intelligibility score across all words.

Table 8.1 – Performance of the speech assessment tasks in the joint analysis, i.e., pathological speech detection and intelligibility assessment using the two considered databases.

Task	English UA-Speech		Dutch COPAS	
Pathological speech detection	Accuracy (%)	AUC	Accuracy (%)	AUC
	96.3	0.98	96.9	0.97
Pathological speech intelligibility assessment	R	R_S	R	R_S
	0.95	0.91	0.84	0.84

8.2.2 Results

In this section, we present the individual performance of the two tasks on our considered databases. In addition, we also provide a brief speaker-wise analysis of the predictions obtained by the two approaches.

Table 8.1 presents the performance of the joint analysis on our considered databases, i.e., the modules for pathological speech detection (in terms of accuracy and AUC) and intelligibility assessment (in terms of Pearson correlation and the Spearman rank correlation). It can be observed that T-GDA yields high pathological speech detection accuracy and AUC performance for both databases. Considering intelligibility assessment for both databases (i.e., HI Dutch-speaking patients in COPAS and CP English-speaking patients in UA-Speech), the P-ESTOI measure yields high and significant correlations with the subjective intelligibility scores.

Figure 8.2 depicts the predicted P-ESTOI intelligibility scores (non-normalized values) and the detection score obtained by the T-GDA approach for Dutch HI patients and healthy speakers in the COPAS database. Figure 8.3 depicts the predicted P-ESTOI intelligibility scores (non-normalized values) and the detection score obtained by T-GDA approach for English CP patients and healthy speakers in the UA-Speech database. The detection score indicates the certainty of the classifier that a given speaker belongs to the healthy class. Considering that the detection threshold is 0.5 (depicted by black dashed horizontal lines in the figures), any speaker with a score higher than 0.5 is then predicted as a healthy speaker and a score lower than 0.5 is predicted as a pathological speaker. Red dashed vertical lines indicate an error made by the detection system, e.g., in Figure 8.2b, only a false positive error has occurred where a healthy speaker (i.e., with index 2) is detected as a patient. Furthermore, in Figure 8.2 it can be observed that the obtained P-ESTOI values are similar across the healthy speakers, and as to be expected they are similar or higher than the scores of the high-intelligible HI patients (i.e., patients with higher predicted intelligibility scores). According to Figure 8.3a, only a false negative error has occurred where a high-intelligible CP patient (i.e., with index 4) is detected as a healthy speaker. It can be also observed that there are more variations in the obtained P-ESTOI values shown in Figure 8.3b across the healthy speakers, while as to be

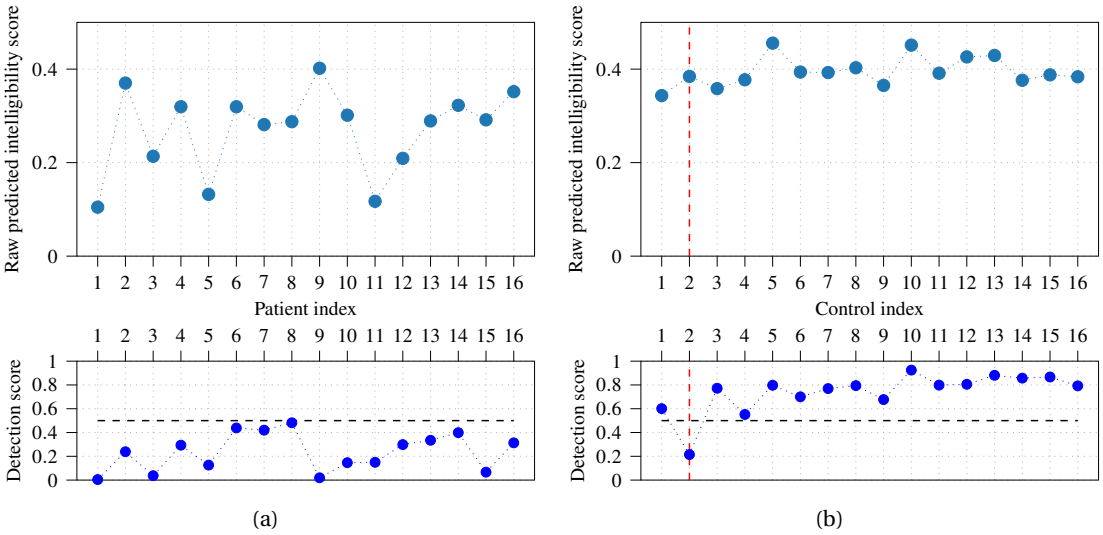


Figure 8.2 – Predicted (non-normalized) P-ESTOI intelligibility scores (upper plot) and predicted speech pathology detection scores (lower plot) for a) HI patients in the COPAS database and for b) healthy speakers in the COPAS database. The detection threshold 0.5 is indicated by the black horizontal dashed line. Red dashed vertical line indicates an error (false positive) made by the detection system, i.e., a healthy speaker is detected as a patient.

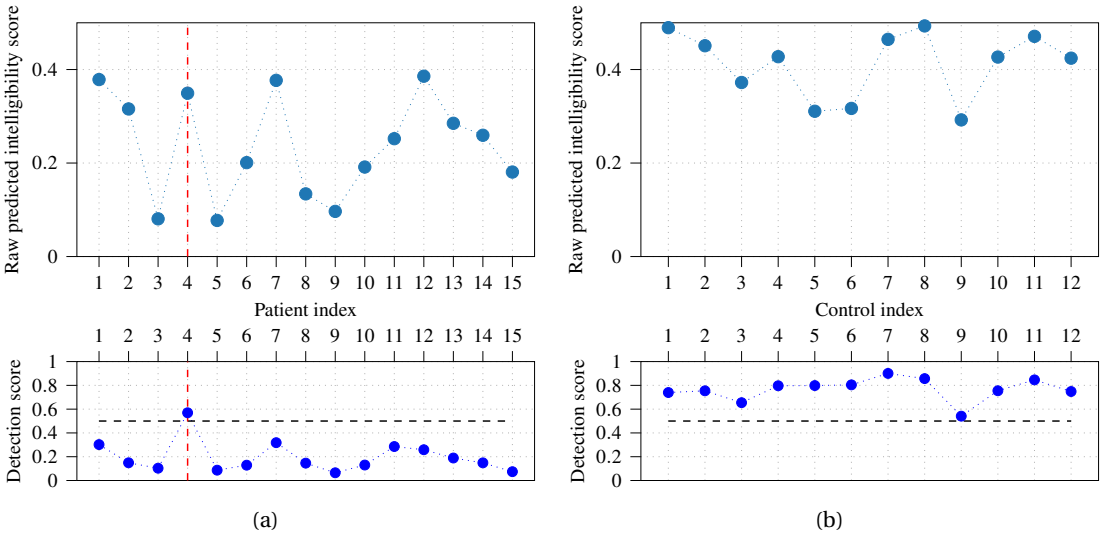


Figure 8.3 – Predicted (non-normalized) P-ESTOI intelligibility scores (upper plot) and predicted speech pathology detection scores (lower plot) for a) CP patients in UA-Speech database and for b) healthy speakers in the UA-Speech database. The detection threshold 0.5 is indicated by the black horizontal dashed line. Red dashed vertical line indicates an error (false negative) made by the detection system, i.e., a high intelligible patient is detected as a healthy speaker.

expected they are similar or higher than the scores of the high-intelligible CP patients (i.e., patients with higher predicted intelligibility scores).

8.3 Summary

In this chapter we have jointly evaluated the two tasks of pathological speech detection and intelligibility assessment, leading to a potential multi-purpose clinical tool useful for clinical practitioners to perform an automatic evaluation of pathological speech. For the speech pathology detection module, we have used and evaluated our proposed temporal subspace-based learning method, while for the intelligibility assessment module, we have used and evaluated our proposed P-ESTOI measure. Experimental results on Dutch and English databases of healthy and pathological speakers have shown that both modules achieve high performance. Although investigating the applicability of such an automatic clinical tool on larger databases with other types of speech disorders is required, we view our contribution here as a step toward designing an automatic assistive tool needed for the pathological speech assessment field.

9 Conclusions and future directions

In this chapter, we summarize the conclusions of this thesis (cf. Section 9.1) and discuss directions of future research (cf. Section 9.2).

9.1 Conclusions

This thesis focused on automatic pathological speech assessment addressing two aspects of automatic acoustic analysis for clinical applications, i.e., pathological speech detection and intelligibility assessment.

We first proposed a subspace-based learning approach to automatically discriminate between pathological and healthy speech. Due to atypical changes in spectro-temporal fluctuations of speech associated with speech disorders (e.g., dysarthria), the dominant spectro-temporal patterns of healthy and pathological speech differ. These patterns are characterized by lower-dimensional subspaces, which are then classified through subspace-based discriminant analysis. Our experimental results have shown that compared to spectral subspaces, temporal subspaces are more successful in discriminating between pathological and healthy speakers and consistently outperform state-of-the-art methods on two databases of different languages and pathologies. However, the temporal subspace-based approach requires time-alignment and having access to utterances with the same phonetic content from both healthy and pathological speakers.

To overcome the need for time-alignment, we investigated the applicability of deep learning as an alternative to classical machine learning-based approaches for pathological speech detection. We proposed novel CNN-based frameworks aiming to learn more robust and relevant features for such a task. First, we explored the feasibility of a pairwise distance-based CNN which relies on comparing two given phonetically-balanced articulatory posterior representations from healthy (reference) and test speakers. The system predicts whether the test representation is from a healthy or pathological speaker after extracting features from input representations and processing the distance matrices computed from them. Experimental

result have shown that our proposed system obtains a good detection performance, is generalisable across languages, and outperforms other baseline CNN-based systems. Such a system does not suffer from the availability of limited training data (usually associated with such tasks) due to the usage of pairwise training. However, similarly to our temporal subspace-based approach, this approach also relies on using utterances with the same phonetic content from both healthy and pathological speakers. In order to investigate the applicability of a CNN-based framework without any phonetic constraints on the speech material from speakers, we proposed supervised speech representation learning frameworks. To reduce the influence of speaker variabilities unrelated to pathology, we proposed to obtain speaker identity-invariant representations by including an adversarial speaker ID auxiliary classifier into the representation training. Further, to obtain a more discriminative representation, we also proposed to supervise the representation learning by a pathological speech auxiliary classifier. Our investigations showed that the proposed representations yield improvement in speech pathology detection performance when compared to unsupervised representation learning frameworks. To avoid using adversarial training for obtaining speaker identity-invariant representations, we also proposed a dual representation learning framework in which separation of the two encoded representations is enforced to either speaker identities cues or cues unrelated to speaker identities. Feature separation is achieved by supervising one of the representations with a speaker ID auxiliary classifier while minimizing a MI criterion between the two representations. Our findings confirm the success of feature separation to obtain speaker-invariant representations without adversarial training, and using the so-obtained speaker-invariant representations improves the pathological speech detection performance compared to unsupervised frameworks. When phonetically-balanced utterances from speakers are available for training, our proposed temporal subspace-based approach for pathological speech detection can be used where it has shown a superior performance compared to other approaches we have proposed in this thesis. When access to such phonetically-balanced utterances is not possible for training, our supervised representation learning approaches can be used instead.

Aiming to automatically assess pathological speech intelligibility, we proposed several non-blind pathological speech intelligibility measures relying on i) creating an intelligible reference representation/model and ii) comparing the reference model to the pathological speech representations/model under evaluation. Our first measure is based on the extended short-time objective intelligibility where utterance-dependent reference representations from multiple healthy speakers are created using a DTW-based clustering method. Intelligibility is then assessed by computing the short-time spectral correlation between the aligned test and reference representations. We showed that this measure is highly correlated with subjective intelligibility ratings for patients with different pathologies outperforming many state-of-the-art pathological speech intelligibility measures while avoiding many of their drawbacks. However, this measure can only be used in scenarios where healthy recordings perfectly matching the phonetic content of the pathological speech signal are available. To increase its flexibility, we also proposed to use synthetic speech generated by state-of-the-art high-quality TTS systems to create intelligible reference representations. We found that the performance

of the intelligibility measure using synthetic speech references is comparable to using natural speech references in phonetically-balanced scenarios, while being also applicable to phonetically-unbalanced scenarios.

To overcome the need for time-alignment, we also proposed a subspace-based intelligibility measure. This measure assesses speech intelligibility by analyzing and comparing the subspaces of healthy and pathological speech through computing a subspace-based distance between them. We have shown that the subspace-based measure can capture pathology-induced distortions in the spectral modulation cues that are important to the perceived speech intelligibility. After proposing two extensions of such a measure, we found that subspace-based measures can outperform several state-of-the-art measures, while being applicable to different scenarios. Considering phonetically-balanced scenarios, our proposed intelligibility measure based on extended short-time objective intelligibility assessment using natural healthy references is shown to be the best performing intelligibility measure among all the measures we have proposed in this thesis. Such a measure can also be used in phonetically-unbalanced scenarios by using synthetic speech references without a substantial decrease in the performance.

Finally, we jointly validated the two tasks of automatic speech pathology detection and speech intelligibility assessment as two components of a multi-purpose clinical tool that can offer objective and automated assistance with pathological speech assessment. Using our temporal subspace-based learning method and the short-time objective intelligibility measure for these tasks, we further confirmed that both modules achieve high performance, independently of the language or disorder.

9.2 Directions for future research

In the following, we provide a few possible directions for future research.

- As mentioned in the thesis, to address the data limitation for pathological speech detection using deep learning, we used two ways to increase the number of training samples, i.e., pairwise word representation training as in our proposed distance-based CNN system and analyzing short (fixed length) segments of speech in our representation learning frameworks. However, as mentioned before, the success of the temporal subspace-based approach in analyzing longer-term acoustic cues for discriminating between pathological and healthy speech might suggest that modeling short segments of speech signals by CNNs may not be sufficiently informative for such a task. In the future, it is worth investigating the applicability of neural architectures that can aggregate information over a longer period of time, while not requiring a large amount of data.
- Although neural architectures have shown promising results for the pathological speech detection task, because of the absence of large training data, they have not yet had a significant dominance over classical machine learning-based approaches. The appli-

cability of transfer learning, i.e., incorporating prior knowledge obtained by training models on another explicit but relevant task with available data, can be investigated.

- Given the availability of enough data, it is worth investigating if our proposed pathological speech detection methods can be extended to automatic classification of different types of speech disorders, which remains an under-explored topic in the literature.
- Due to the difficulty in interpreting high-level feature representations learned by neural networks for pathological speech detection, it is worth investigating the applicability of networks with interpretable convolutional filters (e.g., in SincNet architecture). If such networks are successful in modeling pathological speech, they can also provide interpretable information that can be informative for clinicians.
- To further develop the multi-purpose automatic clinical tool, it is worth investigating training the different tasks jointly by incorporating a multi-task learning framework, as one task might provide useful information for the other task.

Bibliography

A. Dibazar, A., Narayanan, S., and Berger, T. (2002). Feature analysis for automatic detection of pathological speech. In *Proc. IEEE International Conference of Engineering in Medicine and Biology*, volume 1, pages 182 – 183, Houston, TX, USA, USA.

Alexandris, N., Gupta, S., and Koutsias, N. (2017). Remote sensing of burned areas via PCA, part 1; centering, scaling and EVD vs SVD. *Open Geospatial Data, Software and Standards*, 2(1):1–11.

Ali, Z., Elamvazuthi, I., Alsulaiman, M., and Muhammad, G. (2016). Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the all-pole model. *Journal of Voice*, 30(6):757.e7–757.e19.

Allen, J. and Rabiner, L. (1977). A unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564.

Almeida, J. S., Rebouças Filho, P. P., Carneiro, T., Wei, W., Damaševičius, R., Maskeliūnas, R., and de Albuquerque, V. H. C. (2019). Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. *Pattern Recognition Letters*, 125:55–62.

An, K., Kim, M., Teplansky, K., Green, J., Campbell, T., Yunusova, Y., Heitzman, D., and Wang, J. (2018). Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks. In *Proc. 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India.

Ansel, B. M. and Kent, R. D. (1992). Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed Cerebral Palsy. *Journal of Speech and Hearing Research*, 35(2):296–308.

Anumanchipalli, G. K., Meinedo, H., Bugalho, M., Trancoso, I., Oliveira, L. C., and Black, A. W. (2012). Text-dependent pathological voice detection. In *Proc. 13th Annual Conference of the International Speech Communication Association*, pages 530–533, Portland, USA.

Arias-Londoño, J. D., Godino-Llorente, J. I., Sáenz-Lechón, N., Osma-Ruiz, V., and Castellanos-Domínguez, G. (2010). An improved method for voice pathology detection by means of a hmm-based feature space transformation. *Pattern Recogn.*, 43(9):3100–3112.

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. Human Voice and Sounds: From Newborn to Elder.

Baghai-Ravary, L. and Beet, S. (2012). *Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders*. Springer, New York, USA.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. (2018). Mutual information neural estimation. In *Proc. 35th International Conference on Machine Learning*, volume 80, pages 531–540.

Berus, L., Klancnik, S., Brezocnik, M., and Ficko, M. (2019). Classifying Parkinson's disease based on acoustic measures using artificial neural networks. *Sensors (Basel)*, 19(1).

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

Biberger, T. and Ewert, S. D. (2017). The role of short-time intensity and envelope power for speech intelligibility and psychoacoustic masking. *Journal of the Acoustical Society of America*, 142(2):1098–1111.

Bocklet, T., Riedhammer, K., Eysholdt, U., and Haderlein, T. (2012). Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling. *Journal of Voice*, 26(3):390–397.

Bocklet, T., Steidl, S., Noeth, E., and Skodda, S. (2013). Automatic evaluation of Parkinson's speech–acoustic, prosodic and voice related cues. In *Proc. 14th Annual Conference of the International Speech Communication Association*, pages 1149–1153, Lyon, France.

Boersma, P. (2002). PRAAT, a system for doing phonetics by computer. *Glot International*, 5(9):341–345.

Cadima, J. and Jolliffe, I. T. (2009). On relationships between uncentred and column-centred principal component analysis. *Pakistan Journal of Statistics*, 25(4):473–503.

Cardoso, J. and Souloumiac, A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2005). The AMI meeting corpus: A pre-announcement. In *Proc. 2nd International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, pages 28–39, Berlin, Heidelberg. Springer-Verlag.

Castellanos, J., Gómez, S., and Guerra, V. (2002). The triangle method for finding the corner of the L-curve. *Applied Numerical Mathematics*, 43(4):359–373.

Chen, S., Sanderson, C., Harandi, M. T., and Lovell, B. C. (2013). Improved image set classification via joint sparse approximated nearest subspaces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 452–459, Portland, OR, USA.

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. (2020). CLUB: A contrastive log-ratio upper bound of mutual information. In *Proc. 37th International Conference on Machine Learning*, volume abs/2006.12013. PMLR.

Cummins, N., Baird, A., and Schuller, B. W. (2018). Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54. Health Informatics and Translational Data Analytics.

Darley, F. L., Aronson, A. E., and Brown, J. R. (1969). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, 12(2):246–269.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Dejonckere, P. H. and Lebacq, J. (1996). Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology. *Journal for Oto-rhino-laryngology and its Related Specialties*, 58(6):326–332.

Dibazar, A. A., Berger, T. W., and Narayanan, S. S. (2006). Pathological voice assessment. *Proc. IEEE Engineering in Medicine and Biology Society*, 1:1669–1673.

Dimauro, G., Di Nicola, V., Bevilacqua, V., Caivano, D., and Girardi, F. (2017). Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system. *IEEE Access*, 5:22199–22208.

Drullman, R., Festen, J. M., and Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, 95(2):1053–1064.

Dubagunta, S. P. and Magimai-Doss, M. (2019). Using speech production knowledge for raw waveform modelling based Styrian dialect identification. In *Proc. 20th Annual Conference of the International Speech Communication Association*, pages 2383–2387, Graz, Austria.

Duffy, J. R. (2000). *Motor Speech Disorders: Clues to Neurologic Diagnosis, Diagnosis and Treatment Guidelines for the Practicing Physician*, chapter Parkinson's Disease and Movement Disorders, pages 35–53. Humana Press, Totowa, NJ.

Egas-López, J., Orozco, J. R., and Gosztolya, G. (2019). Assessing Parkinson's disease from speech using Fisher vectors. In *Proc. 20th Annual Conference of the International Speech Communication Association*, pages 3063–3067, Graz, Austria.

Elhilali, M., Chi, T., and Shamma, S. A. (2003). A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Communication*, 41(2):331–348.

Elliott, T. M. and Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLOS Computational Biology*, 5(3):1–14.

Enderby, P. (2013). Disorders of communication: Dysarthria. *Handbook of Clinical Neurology*, 110:273–281.

Espinoza-Cuadros, F. M., Perero-Codosero, J. M., Antón-Martín, J., and Hernández-Gómez, L. A. (2020). Speaker de-identification system using autoencoders and adversarial training. *arXiv e-prints*, page arXiv:2011.04696.

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proc. 21st ACM International Conference on Multimedia*, pages 835–838, Barcelona, Spain.

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., Zheng, C., and Chan, W. Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774.

Fletcher, A. R., Wisler, A. A., McAuliffe, M. J., Lansford, K. L., and Liss, J. M. (2017). Predicting intelligibility gains in dysarthria through automated speech feature analysis. *Journal of Speech, Language, and Hearing Research*, 60(11):3058–3068.

Flipsen, P. J. and Parker, R. G. (2008). Phonological patterns in the conversational speech of children with cochlear implants. *Journal of Communication Disorders*, 41(4):337–357.

Fougeron, C., Delvaux, V., Ménard, L., and Laganaro, M. (2018). The MonPaGe_HA database for the documentation of spoken French throughout adulthood. In *Proc. 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *Proc. 32nd International Conference on International Conference on Machine Learning*, volume 37, pages 1180–1189, Lille, France. JMLR.org.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):2096–2030.

García, N., Orozco, J. R., D'Haro, L., Dehak, N., and Noeth, E. (2017). Evaluation of the neurological state of people with Parkinson's disease using i-Vectors. In *Proc. 18th Annual Conference of the International Speech Communication Association*, pages 299–303, Stockholm, Sweden.

Garner, P. (2013). Speech signal processing (SSP) module. https://github.com/idiap/ssp.

Gavidia-Ceballos, L. and Hansen, J. (1996). Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection. *IEEE Transactions on Biomedical Engineering*, 43(4):373–383.

Gillespie, S., Logan, Y.-Y., Moore, E., Laures-Gore, J., Russell, S., and Patel, R. (2017). Cross-database models for the classification of dysarthria presence. In *Proc. 18th Annual Conference of the International Speech Communication Association*, pages 3127–3131, Stockholm, Sweden.

Godino-Llorente, J. I. and Gomez-Vilda, P. (2004). Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Transactions on Biomedical Engineering*, 51(2):380–384.

Godino-Llorente, J. I., Shattuck-Hufnagel, S., Choi, J. Y., Moro-Velázquez, L., and Gómez-García, J. A. (2017a). Towards the identification of idiopathic Parkinson's disease from the speech. new articulatory kinetic biomarkers. *PLoS One*, 12(12):1–35.

Godino-Llorente, J. I., Shattuck-Hufnagel, S., Choi, J. Y., Moro-Velázquez, L., and Gómez-García, J. A. (2017b). Towards the identification of idiopathic Parkinson's disease from the speech. New articulatory kinetic biomarkers. *PLOS ONE*, 12(12):1–35.

Gómez-García, J., Moro-Velázquez, L., and Godino-Llorente, J. (2019). On the design of automatic voice condition analysis systems. part i: Review of concepts and an insight to the state of the art. *Biomedical Signal Processing and Control*, 51:181–199.

Gupta, R., Chaspari, T., Kim, J., Kumar, N., Bone, D., and Narayanan, S. (2016). Pathological speech processing: State-of-the-art, current challenges, and future directions. In *Proc. 41st IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6470–6474.

Haderlein, T., Schützenberger, A., Döllinger, M., and Noeth, E. (2017). Robust automatic evaluation of intelligibility in voice rehabilitation using prosodic analysis. In *Proc. 20th International Conference on Text, Speech, and Dialogue*, pages 11–19, Prague, Czech Republic.

Haderlein, T., Steidl, S., Nöth, E., Rosanowski, F., and Schuster, M. (2004). Automatic recognition and evaluation of tracheoesophageal speech. In *Proc. 7th International Conference on Text, Speech and Dialogue*, pages 331–338, Brno, Czech Republic.

Hamm, J. and Lee, D. D. (2008). Grassmann discriminant analysis: A unifying view on subspace-based learning. In *Proc. 25th International Conference on Machine Learning*, pages 376–383, Helsinki, Finland.

Hansen, P. C. (1992). Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):561–580.

Hegde, S., Shetty, S., Rai, S., and Dodderi, T. (2019). A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, 33(6):947.e11–947.e33.

Hermansky, H. (2011). Speech recognition from spectral dynamics. *Sadhana*, 36(5):729–744.

Higuchi, Y., Tawara, N., Kobayashi, T., and Ogawa, T. (2019). Speaker adversarial training of DPGMM-based feature extractor for zero-resource languages. In *Proc. 20th Annual Conference of the International Speech Communication Association*, pages 266–270, Graz, Austria.

Hinterleitner, F., Norrenbrock, C., and Möller, S. (2013). Is intelligibility still the main problem? A review of perceptual quality dimensions of synthetic speech. In *Proc. 8th International Speech Communication Association Speech Synthesis Workshop*, pages 147–151, Barcelona, Spain.

Hummel, R., Chan, W. Y., and Falk, T. H. (2011). Spectral features for automatic blind intelligibility estimation of spastic dysarthric speech. In *Proc. 12th Annual Conference of the International Speech Communication Association*, pages 3017–3020, Florence, Italy.

Illa, A., Patel, D., Yamini, B. K., Meera, S. S., Shivashankar, N., Veeramani, P.-K., vengalii, S., Polavarapui, K., Nashi, S., Nalini, A., and Ghosh, P. K. (2018). Comparison of speech tasks for automatic classification of patients with Amyotrophic Lateral Sclerosis and healthy subjects. In *Proc. 43rd IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6014–6018, Calgary, Canada.

Imed, L., Waad, B. K., Corinne, F., and Christine, M. (2017). Automatic prediction of speech evaluation metrics for dysarthric speech. In *Proc. 18th Annual Conference of the International Speech Communication Association*, pages 1834–1838, Stockholm, Sweden.

Janbakhshi, P. and Kodrasi, I. (2022). Experimental investigation on stft phase representations for deep learning-based dysarthric speech detection. In *Proc. 47th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Singapore.

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2020). Automatic dysarthric speech detection exploiting subspace-based learning. Idiap-RR Idiap-Internal-RR-12-2020, Idiap.

Janbakhshi, P. and Kodrasi, I. (2021). Supervised speech representation learning for Parkinson's disease classification. In *Proc. 14th ITG Conference on speech communication*, pages 1–5, Virtual Conference.

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2020a). Automatic pathological speech intelligibility assessment exploiting subspace-based analyses. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28(1):1717–1728.

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2020b). Subspace-based learning for automatic dysarthric speech detection. *IEEE Signal Processing Letters*, 28(1):96–100.

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2020c). Synthetic speech references for automatic pathological speech intelligibility assessment. In *Proc. 45th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6099–6103, Virtual Conference.

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2021). Automatic dysarthric speech detection exploiting pairwise distance-based convolutional neural networks. In *Proc. 46th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 7328–7332, Virtual Conference.

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2019a). Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. In *Proc. 44th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 6405–6409, Brighton, UK.

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2019b). Spectral subspace analysis for automatic assessment of pathological speech intelligibility. In *Proc. 20th Annual Conference of the International Speech Communication Association*, pages 3038–3042, Graz, Austria.

Jensen, J. and Taal, C. H. (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022.

Jorgensen, S., Ewert, S. D., and Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *Journal of the Acoustical Society of America*, 134(1):436–446.

Kacha, A., Grenez, F., Orozco-Arroyave, J. R., and Schoentgen, J. (2020). Principal component analysis of the spectrogram of the speech signal: Interpretation and application to dysarthric speech. *Computer Speech & Language*, 59:114–122.

Kalita, S., Mahadeva Prasanna, S. R., and Dandapat, S. (2018). Intelligibility assessment of cleft lip and palate speech using Gaussian posteriograms based on joint spectro-temporal features. *Journal of the Acoustical Society of America*, 144(4):2413–2423.

Karan, B., Sahu, S. S., and Mahto, K. (2020a). Parkinson disease prediction using intrinsic mode function based features from speech signal. *Biocybernetics and Biomedical Engineering*, 40(1):249–264.

Karan, B., Sahu, S. S., and Mahto, K. (2020b). Stacked auto-encoder based time-frequency features of speech signal for Parkinson disease prediction. In *Proc. International Conference on Artificial Intelligence and Signal Processing*, pages 1–4, Amaravati, India.

Karan, B., Sahu, S. S., Orozco-Arroyave, J. R., and Mahto, K. (2020c). Hilbert spectrum analysis for automatic detection and evaluation of Parkinson's speech. *Biomedical Signal Processing and Control*, 61:1–11.

Karan, B., Sahu, S. S., Orozco-Arroyave, J. R., and Mahto, K. (2021). Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction. *Computer Speech & Language*, 69:1–17.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., and Frame, S. (2008). Dysarthric speech database for universal access research. In *Proc. 9th*

Annual Conference of the International Speech Communication Association, pages 1741–1744, Brisbane, Australia.

Kim, J. C., Rao, H., and A Clements, M. (2014). Speech intelligibility estimation using multi-resolution spectral features for speakers undergoing cancer treatment. *Journal of the Acoustical Society of America*, 136(4):315–321.

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.

King, S. and Karaiskos, V. (2016). The Blizzard challenge 2016. In *Proc. Blizzard Challenge Workshop*, Cupertino, USA.

King, S., Wihlborg, L., and Guo, W. (2017). The Blizzard challenge 2017. In *Proc. Blizzard Challenge Workshop*, Stockholm, Sweden.

Kingma, D. P. and Ba, J. (2015). ADAM: a method for stochastic optimization. In *Proc. 3rd International Conference on Learning Representations*, San Diego, CA, USA.

Kodrasi, I. and Bourlard, H. (2019). Super-Gaussianity of speech spectral coefficients as a potential biomarker for dysarthric speech detection. In *Proc. 44th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6400–6404, Brighton, UK.

Kodrasi, I. and Bourlard, H. (2020). Spectro-temporal sparsity characterization for dysarthric speech detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28(1):1210–1222.

Kodrasi, I., Goetze, S., and Doclo, S. (2013). Regularization for partial multichannel equalization for speech dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 21(9):1879–1890.

Kominek, J. and Black, A. (2004). The CMU Arctic speech databases. In *Proc. 5th International Speech Communication Association Speech Synthesis Workshop*, pages 223–224, Pittsburgh, USA.

Korzekwa, D., Barra-Chicote, R., Kostek, B., Drugman, T., and Lajszczak, M. (2019). Interpretable deep learning model for the detection and reconstruction of dysarthric speech. In *Proc. 20th Annual Conference of the International Speech Communication Association*, pages 3890–3894, Austria, Graz.

Ku, W., Storer, R. H., and Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 30(1):179–196.

Kwatra, V. and Han, M. (2010). Fast covariance computation and dimensionality reduction for sub-window features in images. In *Computer Vision – ECCV 2010*, pages 156–169, Berlin, Heidelberg.

Landa, S., Pennington, L., Miller, N., Robson, S., Thompson, V., and Steen, N. (2014). Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding. *International Journal of Speech-Language Pathology*, 16(4):408–416.

Le, L., Patterson, A., and White, M. (2018). Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In *Proc. International Conference on Neural Information Processing Systems*, pages 107–117, Montréal, Canada.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Li, H., Tu, M., Huang, J., Narayanan, S., and Georgiou, P. (2020). Speaker-invariant affective representation learning via adversarial training. In *Proc. 45th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7144–7148, Barcelona, Spain.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., and Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4):1015–1022.

Luo, G., Wei, J., Hu, W., and Maybank, S. J. (2019). Tangent Fisher vector on matrix manifolds for action recognition. *IEEE Transactions on Image Processing*, 29:3052–3064.

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., and Nöth, E. (2009). PEAKS - A system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5):425–437.

Makhoul, J. and Cosell, L. (1976). LPCW: An LPC vocoder with linear predictive spectral warping. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 466–469, Philadelphia, PA, USA.

Mallela, J., Illa, A., Belur, Y., Atchayaram, N., Yadav, R., Reddy, P., Gope, D., and Ghosh, P. K. (2020). Raw speech waveform based classification of patients with ALS, Parkinson's disease and healthy controls using CNN-BLSTM. In *Proc. 21st Annual Conference of the International Speech Communication Association*, pages 4586–4590, Shanghai, China.

Marelli, F., Schnell, B., Bourlard, H., Dutoit, T., and Garner, P. N. (2019). An end-to-end network to synthesize intonation using a generalized command response model. In *Proc. 44th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 7040–7044, Brighton, UK.

Martínez, D., Green, P., and Christensen, H. (2013). Dysarthria intelligibility assessment in a factor analysis total variability space. In *Proc. 14th Annual Conference of the International Speech Communication Association*, pages 2133–2137, Lyon, France.

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.

Meinard, M. (2007). *Information Retrieval for Music and Motion*, chapter Dynamic Time Warping, pages 69–84. Springer, Berlin, Heidelberg.

Meng, Z., Li, J., Chen, Z., Zhao, Y., Mazalov, V., Gong, Y., and Juang, B.-H. (2018). Speaker-invariant training via adversarial learning. In *Proc. 43rd IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5969–5973, Calgary, Canada.

Michaelis, D., Frohlich, M., and Strube, H. W. (1998). Selection and combination of acoustic features for the description of pathologic voices. *Journal of the Acoustical Society of America*, 103(3):1628–1639.

Middag, C., Martens, J.-P., Nuffelen, G. V., 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.

Middag, C., Saeys, Y., and Martens, J.-P. (2010). Towards an ASR-free objective analysis of pathological speech. In *Proc. 11th Annual Conference of the International Speech Communication Association*.

Middag, C., Van Nuffelen, G., Martens, J. P., and De Bodt, M. (2008). Objective intelligibility assessment of pathological speakers. In *Proc. 9th Annual Conference of the International Speech Communication Association*, pages 1745–1748.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., and Mullers, K. R. (1999). Fisher discriminant analysis with kernels. In *Proc. 1999 IEEE Signal Processing Society Workshop*, pages 41–48, Madison, WI, USA.

Mishra, B., Kasai, H., Jawanpuria, P., and Saroop, A. (2019). A Riemannian gossip approach to subspace learning on Grassmann manifold. *Machine Learning*, 108(10):1783–1803.

Moro-Velázquez, L., Gómez-García, J. A., Godino-Llorente, J. I., Villalba, J., Orozco-Arroyave, J. R., and Dehak, N. (2018). Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's disease. *Applied Soft Computing*, 62:649–666.

Moyer, D., Gao, S., Brekelmans, R., Steeg, G. V., and Galstyan, A. (2018). Invariant representations without adversarial training. In *Proc. 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 9102–9111, Red Hook, NY, USA. Curran Associates Inc.

Norel, R., Pietrowicz, M., Agurto, C., Rishoni, S., and Cecchi, G. (2018). Detection of Amyotrophic Lateral Sclerosis (ALS) via acoustic analysis. In *Proc. 19th Annual Conference of the International Speech Communication Association*, pages 377–381, Hyderabad, India.

Nuffelen, G. V., De Bodt, M., Middag, C., and Martens, J.-P. (2009a). Dutch corpus of pathological and normal speech (COPAS). Technical report, Antwerp University Hospital and Ghent University, Belgium.

Nuffelen, G. V., Middag, C., De Bodt, M., and Martens, J.-P. (2009b). Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language & Communication Disorders*, 44(5):716–730.

Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. *Folia Phoniatrica et Logopaedica*, 61(1):49–56.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

Orozco-Arroyave, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J., González-Rátiva, M., and Noeth, E. (2014a). New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *Proc. 9th International Conference on Language Resources and Evaluation*, pages 342–347, Reykjavik, Iceland.

Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Bonilla, J. F. V., Skodda, S., Rusz, J., and Nöth, E. (2014b). Automatic detection of Parkinson's disease from words uttered in three different languages. In *Proc. 15th Annual Conference of the International Speech Communication Association*, Singapore.

Orozco-Arroyave, J. R., Honig, F., Arias-Londono, J. D., Vargas-Bonilla, J. F., Daqrouq, K., Skodda, S., Rusz, J., and Noeth, E. (2016a). Automatic detection of Parkinson's disease in running speech spoken in three different languages. *The Journal of the Acoustical Society of America*, 139(1):481–500.

Orozco-Arroyave, J. R., Honig, F., Arias-Londono, J. D., Vargas-Bonilla, J. F., Daqrouq, K., Skodda, S., Rusz, J., and Noth, E. (2016b). Automatic detection of Parkinson's disease in running speech spoken in three different languages. *Journal of the Acoustical Society of America*, 139(1):481–500.

Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., and Noeth, E. (2015a). Spectral and cepstral analyses for Parkinson's disease detection in Spanish vowels and words. *Expert Systems*, 32(6):688–697.

Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Skodda, S., Rusz, J., and Nöth, E. (2015b). Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease. In *Proc. 16th Annual Conference of the International Speech Communication Association*, pages 95–99, Dresden, Germany.

Orozco-Arroyave, J. R., Vargas-Bonilla, J. F., and Delgado-Trejos, E. (2012). Acoustic analysis and non linear dynamics applied to voice pathology detection: A review. *Recent Patents on Signal Processing*, 2(12):96–107.

Paja, M. S. and Falk, T. H. (2012). Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech. In *Proc. 13th Annual Conference of the International Speech Communication Association*, pages 62–65, Oregon, USA.

Parsa, V. and Jamieson, D. G. (2001). Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. *Journal of Speech, Language, and Hearing Research*, 44(2):327–339.

Pasley, B., Flinker, A., and Knight, R. (2015). Speech sounds. In *Brain Mapping*, pages 661–666. Academic Press, Waltham.

Peng, X., Huang, Z., Sun, X., and Saenko, K. (2019). Domain agnostic learning with disentangled representations. In *Proc. 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5102–5112, California, USA. PMLR.

Rabiner, L. and Juang, B. (1993). *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs.

Ram, D., Miculicich, L., and Bourlard, H. (2020). Neural network based end-to-end query by example spoken term detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28(1):1416–1427.

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

Rosen, K. M., Kent, R. D., Delaney, A. L., and Duffy, J. R. (2006). Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers. *Journal of Speech Language and Hearing Research*, 49(2):395–411.

Ruiping Wang, Shiguang Shan, Xilin Chen, and Wen Gao (2008). Manifold-manifold distance with application to face recognition based on image set. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, AK, USA.

Schnell, B. and Garner, P. N. (2018). A neural model to predict parameters for a generalized command response model of intonation. In *Proc. 19th Annual Conference of the International Speech Communication Association*, pages 3147–3151, Hyderabad, India.

Schölkopf, B., Smola, A. J., and Müller, K.-R. (1997). Kernel principal component analysis. In *Proc. 7th International Conference on Artificial Neural Networks*, ICANN '97, pages 583–588, Berlin, Heidelberg. Springer-Verlag.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. 14th Annual Conference of the International Speech Communication Association*, pages 148–152, Lyon, France.

Schuster, M., Nöth, E., Haderlein, T., Steidl, S., Batliner, A., and Rosanowski, F. (2005). Can you understand him? Let's look at his word accuracy-automatic evaluation of tracheoesophageal speech. In *Proc. 30th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA.

Sha, L. and Lukasiewicz, T. (2021). Multi-type disentanglement without adversarial training. In *Proc. 35th AAAI Conference on Artificial Intelligence*, pages 9515–9523, Virtual Conference.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304.

Soldo, S., Magimai-Doss, M., and Bourlard, H. (2012). Synthetic references for template-based ASR using posterior features. In *Proc. 13th Annual Conference of the International Speech Communication Association*, pages 52–57, Portland, USA.

Steeneken, H. J. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251.

Sussman, J. E. and Tjaden, K. (2012). Perceptual measures of speech from individuals with Parkinson's disease and multiple Sclerosis: intelligibility and beyond. *Journal of Speech, Language, and Hearing Research*, 55(4):1208–1219.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proc. 35th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4214–4217, Dallas, TX, USA.

Tammen, M., Kodrasi, I., and Doclo, S. (2018). Complexity reduction of eigenvalue decomposition-based diffuse power spectral density estimators using the power method. In *Proc. 43rd IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 451–455, Calgary, AB, Canada.

Travieso, C. M., Alonso, J. B., Orozco-Arroyave, J., Vargas-Bonilla, J., Nöth, E., and Ravelo-García, A. G. (2017). Detection of different voice diseases based on the nonlinear characterization of speech signals. *Expert Systems with Applications*, 82:184–195.

Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., and Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 59(5):1264–1271.

Vaiciukynas, E., Gelzinis, A., Verikas, A., and Bacauskiene, M. (2017a). Parkinson's disease detection from speech using convolutional neural networks. In *Proc. International Conference on Smart Objects and Technologies for Social Good*, pages 206–215, Pisa, Italy. Springer International Publishing.

Vaiciukynas, E., Verikas, A., Gelzinis, A., and Bacauskiene, M. (2017b). Detecting Parkinson's disease from sustained phonation and speech signals. *PLOS ONE*, 12(10):1–16.

Van Der Veen, A. J., Deprettere, E. F., and Swindlehurst, A. L. (1993). Subspace-based signal analysis using singular value decomposition. *Proceedings of the IEEE*, 81(9):1277–1308.

Vasquez, J., Orozco, J. R., and Noeth, E. (2017). Convolutional neural network to model articulation impairments in patients with Parkinson's disease. In *Proc. 18th Annual Conference of the International Speech Communication Association*, pages 314–318, Stockholm, Sweden.

Vasquez-Correa, J., Arias-Vergara, T., Schuster, M., Orozco-Arroyave, J., 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.

Wall, M. E., Rechtsteiner, A., and Rocha, L. M. (2003). *A Practical Approach to Microarray Data Analysis*, chapter Singular Value Decomposition and Principal Component Analysis, pages 91–109. Springer US, Boston, MA.

Wallen, E. J. and Hansen, J. H. L. (1996). A screening test for speech pathology assessment using objective quality measures. In *Proc. 4th International Conference on Spoken Language Processing*, volume 2, pages 776–779 vol.2, Philadelphia, PA, USA.

Wang, D., Deng, L., Yeung, Y. T., Chen, X., Liu, X., and Meng, H. (2021). Unsupervised domain adaptation for dysarthric speech detection via domain adversarial training and mutual information minimization. In *Proc. 22nd Annual Conference of the International Speech Communication Association*, pages 2956–2960, Brno, Czechia.

Wang, J., Kothalkar, P., Cao, B., and Heitzman, D. (2016). Towards automatic detection of Amyotrophic Lateral Sclerosis from speech acoustic and articulatory samples. In *Proc. 17th Annual Conference of the International Speech Communication Association*, pages 1195–1199, San Francisco, USA.

Wang, T. and Shi, P. (2009). Kernel grassmannian distances and discriminant analysis for face recognition from image sets. *Pattern Recognition Letters*, 30(13):1161–1165.

Windrich, M., Maier, A., Kohler, R., Nöth, E., Nkenke, E., Eysholdt, U., and Schuster, M. (2008). Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. *Folia Phoniatrica et Logopaedica*, 60(3):151–156.

Wu, Z., Watts, O., and King, S. (2016). Merlin: An open source neural network speech synthesis system. In *Proc. 9th International Speech Communication Association Speech Synthesis Workshop*, pages 202–207, Sunnyvale, USA.

Xin, D., Komatsu, T., Takamichi, S., and Saruwatari, H. (2021). Disentangled speaker and language representations using mutual information minimization and domain adaptation for cross-lingual TTS. In *Proc. 46th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6608–6612, Virtual Conference.

yao Hu, T., Shrivastava, A., Tuzel, O., and Dhir, C. S. (2020). Unsupervised style and content separation by minimizing mutual information for speech synthesis. In *Proc. 45th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3267–3271, Virtual Conference.

Ye, K. and Lim, L. H. (2016). Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197.

Yuan, S., Cheng, P., Zhang, R., Hao, W., Gan, Z., and Carin, L. (2021). Improving zero-shot voice style transfer via disentangled representation learning. In *Proc. International Conference on Learning Representations*, pages 1–12, Vienna, Austria.

Zeng, F.-G., Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (2005). Speech recognition with amplitude and frequency modulations. *Proceedings of the National Academy of Sciences*, 102(7):2293–2298.

Zhao, Z. and Liu, F. (2004). Industrial monitoring based on moving average pca and neural network. In *Proc. 30th Annual Conference of IEEE Industrial Electronics Society*, volume 3, pages 2168–2171, Busan, South Korea.

Zwillinger, D. and Kokoska, S. (2000a). *CRC Standard Probability and Statistics Tables and Formula*, chapter Nonparametric Statistics. Chapman and Hall, New York.

Zwillinger, D. and Kokoska, S. (2000b). *CRC Standard Probability and Statistics Tables and Formula*, chapter Standard Normal Distribution. Chapman and Hall, New York.

Parvaneh Janbakhshi

email: pr.janbakhshi@gmail.com
Webpage: pjanbakhshi.github.io

Research Interests

Speech and Audio Signal Processing (for clinical applications), Machine Learning and Deep Learning, Biological Signal Processing

Education

2018–2022 Doctor of Philosophy (PhD) in Electrical Engineering
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
GPA: 5.5/6, 12 credits

2014–2016 Master of Science in Biomedical Engineering (Bioelecrtics)
Sharif University of Technology, Tehran, Iran
GPA: 4/4 (18.88/20), 29 credits–ranked 2^{nd}

2009–2014 Bachelor of Science in Biomedical Engineering (Bioelecrtics)
Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran
GPA: 3.70/4 (17.27/20), 140 credits

Research Experience

Doctoral Researcher in Idiap Research Institute, Martigny, Switzerland
- Thesis: Automatic pathological speech assessment
  - Supervisor: Prof. H. Bourlard and Dr. I. Kodarsi, 2022
- Relevant coursework: Deep Learning, Optimization for Machine Learning, Statistical Sequence Processing

Researcher of Cognitive Neurobiology Laboratory in School of Cognitive Sciences Institute, Tehran, Iran:
- Project: Investigating the phase amplitude coupling in the middle temporal visual area of rhesus monkeys
  - Supervisors: Dr. M. R. Daliri and Dr. M. Esghaei, 2017

Master of Science Thesis in Sharif University of Technology, Tehran, Iran:
- Thesis: Extraction of respiratory information from ECG and its application for sleep apnea detection
  - Supervisor: Prof. M. B. Shamsollahi, 2016
- Relevant coursework: Pattern Recognition, Biological Signal Processing

Bachelor of Science Thesis in Amirkabir University of Technology, Tehran, Iran:
- Thesis: Designing and implementing an automatic neuromuscular electro-stimulation device to prevent diseases such as deep vein thrombosis and varicose veins
  - Supervisor: Dr. A. Maleki, 2014

Publications

- **Janbakhshi, P.**, Kodrasi, I., "Adversarial-free speaker identity-invariant representation rearning for automatic dysarthric speech classification", in Proc. INTERSPEECH, 2022 (under review).
- **Janbakhshi, P.**, Kodrasi, I., "Experimental investigation on STFT phase representations for deep learning-based dysarthric speech detection", in Proc. ICASSP, 2022.
- **Janbakhshi, P.**, Kodrasi, I., "Supervised speech representation learning for Parkinson's disease classification", in Proc. ITG Conference on Speech Communication, 2021.
- **Janbakhshi, P.**, Kodrasi, I., Bourlard, H., "Automatic dysarthric speech detection exploiting pairwise distance-based convolutional neural networks", in Proc. ICASSP, 2021.
- **Janbakhshi, P.**, Kodrasi, I., Bourlard, H., "Subspace-based learning for automatic dysarthric speech detection", Signal Processing Letters, 2021.
- **Janbakhshi, P.**, Kodrasi, I., Bourlard, H., "Automatic pathological speech intelligibility assessment exploiting subspace-based analyses," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020.
- **Janbakhshi, P.**, Kodrasi, I., Bourlard, H., "Synthetic speech references for automatic pathological speech intelligibility assessment", in Proc. ICASSP, 2020.
- **Janbakhshi, P.**, Kodrasi, I., Bourlard, H., "Spectral subspace analysis for automatic assessment of pathological speech intelligibility,", in Proc. INTERSPEECH, 2019.
- **Janbakhshi, P.**, Kodrasi, I., Bourlard, H., "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure", in Proc. ICASSP, 2019.
- **Janbakhshi, P.**, Shamsollahi, M. B., "Sleep apnea detection from single-lead ECG using features based on ECG-derived respiration (EDR) signals", IRBM, 2018.
- **Janbakhshi, P.**, Shamsollahi, M. B., "ECG-derived respiration estimation from single-lead ECG using gaussian process and phase space reconstruction methods", Biomedical Signal Processing and Control, 2018.
- Maleki, A., **Janbakhshi, P.**, "Intelligent device for preventing varicose and deep vein thrombosis based on electrical stimulation", Patented in Iran, Patent No. 83492, 2014

Honors & Awards

- PhD student award by Idiap Research Institute
- **Ranked 2** in Master of Science, Bioelectric Major, Electrical Engineering Department, Sharif University of Technology (2016)
- **Ranked 50** among more than 15000 participants in Nationwide University Entrance Exam in Master of Science, Biomedical Engineering (2014)
- Bachelor of Science thesis was awarded by the university as the **best BSc project** of the year in Bioelectric Engineering.

Ad-Hoc Reviewer

Nature Scientific Reports, Springer Behavior Research Methods, Elsevier Computers in Biology and Medicine, Elsevier Speech Communication, IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Computer Skills

Technical Softwares:	Pytorch, Matlab, Praat
Programming languages:	Python