EPFL

Thèse n° 10 679

Towards Integrated Processing of Physiological Signals and Speech

Présentée le 28 mars 2025

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

Zohreh MOSTAANI

Acceptée sur proposition du jury

Prof. D. N. A. Van De Ville, président du jury
Prof. D. Gatica-Perez, Dr M. Magimai Doss, directeurs de thèse
Dr V. Mitra, rapporteur
Dr M. Cernak, rapporteur
Prof. J.-Ph. Thiran, rapporteur

École polytechnique fédérale de Lausanne

2025

To my family…

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Mathew Magimai Doss, for his unwavering support and guidance throughout my Ph.D. journey. His knowledge, encouragement, and patience have been invaluable to my development as a researcher. He always answered my questions with thoroughness, and his insightful advice has helped shape my work in profound ways. For this, I am truly grateful.

I would also like to thank my thesis director, Prof. Daniel Gatica-Perez, for his thoughtful comments and constructive suggestions. My appreciation extends to the jury members, Prof. Jean-Philip Thiran, Dr. Vikramjit Mitra, and Dr. Milos Cernak, for their time and insightful feedback, as well as the jury president, Prof. Dimitri Van De Ville. I am also thankful to the administrative staff at Idiap, especially Sylvie Meier and Laura Coppey, for their assistance and support throughout this process.

This thesis is the result of collaboration and interactions with many colleagues and peers who have enriched my experience. I am grateful to S. Pavankumar Dubagunta for his early guidance at the beginning of my journey, and to Venkata Srikanth Nallanthighal from Philips Research in the Netherlands for collaborating on essential groundwork for my thesis. My sincere thanks also go to Gürkan Yilmaz from the Swiss Center for Electronics and Microtechnology (CSEM) for providing the device used for data collection in my research. Additionally, I am grateful to Vera Lehmann from the Department of Diabetes, Endocrinology, Nutritional Medicine, and Metabolism (UDEM) at Inselspital, Bern University Hospital, University of Bern, for designing the clinical study featured in this thesis. I extend my thanks to Colombine Verzat from the Idiap development team for her support and for providing the software needed for data collection. My thanks also go to Gasser Elbanna for his valuable contributions during his internship at Idiap, resulting in a publication that is part of this thesis. I would also like to thank my co-authors throughout my Ph.D. studies for their valuable insights and collaboration. My appreciation extends to the IT team at Idiap for their technical support. I also acknowledge that ChatGPT, a generative artificial intelligence chatbot developed by OpenAI, has been used in parts of the thesis to increase readability. All the analyses, interpretations, and conclusions are my own.

This research was made possible by the funding from the Swiss National Science Foundation (SNSF) through the project Towards Integrated processing of Physiological and Speech signals

(TIPS), grant no. 200021_188754, which provided me with the financial support needed to conduct my study and attend various conferences where I had the opportunity to present my work, receive valuable feedback, and connect with the research community.

The Ph.D. journey can be long and full of challenges, but the friendships I've made along the way have truly been the highlight. Sargam, Apoorv, Parvaneh, Weipang, Suhan, Pablo, Neha, Laurent, Florian, Eklavya, Tilak, François, Sarthak, Angel, Julian, Enno, Fabio, Andrei, and so many others—thank you for being there. From hikes and coffee breaks to cooking nights and parties, each moment together made this journey more enjoyable and gave me the strength to push through the hard times.

I'm also so grateful for the support of my family, which has been a constant through everything. My deepest thanks go to my parents, who have always believed in me and encouraged my dreams. And to my brothers, whose unwavering support has been my steady anchor—I can't thank you enough. I'm also incredibly thankful to my mother-in-law for her kindness and support during the tougher times. Having all of you behind me has meant the world.

Last but certainly not least, I want to share my deepest gratitude for my husband, Amir. Your unwavering love and support mean everything to me. You've encouraged me to keep pushing forward, always reminding me of my potential, and you've been my rock when I needed it most. I'm endlessly grateful for our beautiful son and the family we have built together—There is nothing I cherish more than the love and strength we share. Thank you for everything, always.

Martigny, January 2, 2025 Zohreh

Abstract

Research in speech processing has largely focused on source-system modeling, where vocal fold vibrations serve as the source and the vocal cavity's articulations as the system. Nonetheless, speech production includes a complex combination of physiological systems including muscular, respiratory, cognitive, and nervous systems. Variations in these systems can significantly affect speech. For example, individuals with respiratory or cardiovascular issues may experience breathlessness that alters their speech. Parkinson's Disease (PD), a neurodegenerative disorder, can impair speech by disrupting the muscle control required for articulation. Variations in cognitive load from mental stress can also impair speech capabilities. A deeper understanding of the relationship between speech and physiological signals could enhance existing speech technologies and lead to new applications particularly in the healthcare domain. In this thesis we move beyond the traditional speech processing methods and investigate physiological signals in relation to speech. More specifically, we estimate breathing patterns and heart rate from speech signals and integrate them into speech related applications.

We developed end-to-end convolutional neural networks to estimate breathing patterns from raw waveform speech signals and compared them with models using spectral features. The evaluation employed standard regression metrics and breathing related parameters, such as breathing rate, and tidal volume. We showed that both models performed similarly, with raw waveform models requiring a smaller input window. Our single and cross database analyses confirmed the generalizability of the models. We also examined the limitations of the evaluation metrics employed in our study. Additionally, we analysed the raw waveform based models to understand the information they model. Our experiments revealed that they rely on the low-frequency components of the speech signals for accurate estimation of breathing patterns. Furthermore, we studied neural embeddings extracted from the raw waveform based models in various applications, including COVID-19 detection from speech, emotion recognition, and analysing breathing information differences in natural versus synthetic speech for presentation attack detection.

We also created models to estimate cardiac parameters like heart rate from speech using acoustic features and neural embeddings derived from self-supervised learning models. We found significant speaker dependent variability in performance. Additionally, our approach was validated on two datasets, producing consistent trends and confirming model generalizability.

Abstract

Finally, we studied the feasibility of applying the developed methodologies in a clinical setting by detecting hypoglycemic states in diabetic patients through speech analysis. For this, we employed neural embeddings from breathing pattern estimation networks alongside other neural embeddings and acoustic features. We also examined the performance of heart rate estimation models in this context. As part of this research, we compiled two novel datasets with simultaneous recordings of speech and physiological signals, one of which was collected in a clinical environment. These datasets were used to evaluate the performance of the developed models.

Keywords: Speech processing, breathing pattern estimation, breathing parameters, heart rate, physiological signals, machine learning, convolutional neural networks, neural embeddings, COVID-19 detection, hypoglycemia.

Zusammenfassung

Forschung in der Sprachverarbeitung hat sich weitgehend auf das Quelle-Filter-Modell konzentriert, bei dem die Schwingungen der Stimmbänder als Quelle und die Artikulationen im Vokaltrakt als Filter dienen. Dennoch umfasst die Sprachproduktion eine komplexe Kombination physiologischer Systeme, einschließlich Muskel-, Atem- und Nervensystem. Veränderungen in diesen Systemen können die Sprache erheblich beeinflussen. Beispielsweise können Personen mit respiratorischen oder kardiovaskulären Problemen unter Atemnot leiden, die ihre Sprache verändert. Parkinson-Krankheit (PD), eine neurodegenerative Erkrankung, kann die Sprache beeinträchtigen, indem sie die für die Artikulation erforderliche Muskelkontrolle stört. Auch Variationen der kognitiven Belastung durch mentalen Stress können die Sprachfähigkeiten beeinträchtigen. Ein tieferes Verständnis der Beziehung zwischen Sprache und physiologischen Signalen könnte bestehende Sprachtechnologien verbessern und zu neuen Anwendungen insbesondere im Gesundheitsbereich führen. In dieser Arbeit gehen wir über traditionelle Methoden der Sprachverarbeitung hinaus und untersuchen physiologische Signale im Zusammenhang mit Sprache. Insbesondere schätzen wir Atemmuster und Herzfrequenz anhand von Sprachsignalen und integrieren diese in sprachbezogene Anwendungen.

Wir haben Ende-zu-Ende-Konvolutionsnetze entwickelt, um Atemmuster aus rohen Sprachwellenformen zu ermitteln, und diese mit Modellen verglichen, die spektrale Merkmale verwenden. Die Auswertung erfolgte anhand standardmäßiger Regressionsmetriken sowie atembezogener Parameter wie Atemfrequenz und Atemzugvolumen. Wir zeigten, dass beide Modelle ähnlich leistungsfähig sind, wobei Modelle mit rohen Wellenformen ein kleineres Fenster benötigen. Unsere einzelnen und datenbankübergreifenden Analysen bestätigten die Generalisierbarkeit der Modelle. Zudem untersuchten wir die Beschränkungen der in unserer Studie eingesetzten Bewertungsmetriken. Darüber hinaus analysierten wir die Rohwellenmodelle, um zu verstehen, welche Informationen sie modellieren. Unsere Experimente zeigten, dass sie sich auf die niederfrequenten Komponenten der Sprachsignale stützen, um Atemmuster präzise zu schätzen. Ferner untersuchten wir die neuronalen Einbettungen der Rohwellenmodelle in verschiedenen Anwendungen, einschließlich der COVID-19-Erkennung anhand von Sprache, Emotionserkennung und der Analyse von Ateminformationsunterschieden in natürlicher und synthetischer Sprache zur Erkennung von Präsentationsangriffen.

Wir entwickelten auch Modelle, die aus Sprache kardiovaskuläre Parameter wie die Herzfrequenz schätzen, basierend auf akustischen Merkmalen und neuronalen Einbettungen, die

aus selbstüberwachten Lernmodellen abgeleitet wurden. Dabei stellten wir eine signifikante sprecherabhängige Variabilität in der Leistung fest. Unser Ansatz wurde auf zwei Datensätzen validiert, wobei konsistente Trends die Generalisierbarkeit der Modelle bestätigten.

Abschließend untersuchten wir die Umsetzbarkeit der entwickelten Methodologien in einer klinischen Umgebung, indem wir hypoglykämische Zustände bei Diabetespatienten durch Sprachanalyse feststellen. Dafür verwendeten wir neuronale Einbettungen aus Netzwerken zur Atemmustervorhersage sowie andere neuronale Einbettungen und akustische Merkmale. Zudem prüften wir die Leistung der Modelle zur Schätzung der Herzfrequenz in diesem Kontext. Im Rahmen dieser Forschung erstellten wir zwei neuartige Datensätze mit gleichzeitigen Aufzeichnungen von Sprache und physiologischen Signalen, von denen einer in einer klinischen Umgebung gesammelt wurde. Diese Datensätze wurden zur Evaluierung der entwickelten Modelle verwendet.

Schlüsselwörter: Sprachverarbeitung, Atemmuster-Schätzung, Atemparameter, Herzfrequenz, physiologische Signale, maschinelles Lernen, Convolutional Neural Networks, neuronale Einbettungen, COVID-19-Erkennung, Hypoglykämie.

Résumé

La recherche sur le traitement de la parole s'est principalement axée sur la modélisation source-système, où les vibrations des cordes vocales servent de source et les articulations de la cavité vocale comme système. Néanmoins, la production de la parole inclut une combinaison complexe de systèmes physiologiques, notamment les systèmes musculaire, respiratoire, cognitif et nerveux. Les variations au sein de ces systèmes peuvent avoir un impact significatif sur la parole. Par exemple, les individus présentant des problèmes respiratoires ou cardiovasculaires peuvent subir un essoufflement qui altère leur discours. La maladie de Parkinson (MP), un trouble neurodégénératif, peut affecter la parole en perturbant le contrôle musculaire nécessaire à l'articulation. De plus, des variations de la charge cognitive dues au stress mental peuvent également altérer les capacités verbales. Une compréhension plus approfondie de la relation entre la parole et les signaux physiologiques pourrait améliorer les technologies de traitement de la parole existantes et mener à de nouvelles applications, notamment dans le domaine de la santé. Dans cette thèse, nous allons au-delà des méthodes traditionnelles de traitement de la parole et nous étudions les signaux physiologiques en relation avec la parole. Plus précisément, nous estimons les schémas respiratoires et la fréquence cardiaque à partir de signaux vocaux et les intégrons dans des applications liées à la parole.

Nous avons développé des réseaux neuronaux convolutifs de bout en bout pour estimer les schémas respiratoires à partir de signaux vocaux bruts (ondes sonores) et les avons comparés à des modèles utilisant des caractéristiques spectrales. L'évaluation a utilisé des métriques de régression standard ainsi que des paramètres liés à la respiration, tels que la fréquence respiratoire et le volume courant. Nous avons démontré que les deux types de modèles offrent des performances similaires, les modèles utilisant l'onde sonore brute nécessitant une fenêtre d'entrée plus petite. Nos analyses, à la fois sur des bases de données uniques et croisées, ont confirmé la généralisation des modèles. Nous avons également examiné les limites des métriques d'évaluation employées dans notre étude. De plus, nous avons analysé les modèles basés sur l'onde sonore brute afin de comprendre les informations qu'ils modélisent. Nos expériences ont révélé qu'ils s'appuient sur les composantes basses fréquences des signaux vocaux pour estimer avec précision les schémas respiratoires. Par ailleurs, nous avons étudié les représentations neuronales issues des modèles basés sur l'onde sonore brute dans diverses applications, notamment la détection du COVID-19 par la parole, la reconnaissance des émotions et l'analyse des différences d'informations respiratoires dans la parole naturelle par

rapport à la parole synthétique pour la détection des attaques de présentation.

Nous avons également conçu des modèles pour estimer des paramètres cardiaques tels que la fréquence cardiaque à partir de la parole, en utilisant des caractéristiques acoustiques et des représentations neuronales dérivées de modèles d'apprentissage auto-supervisé. Nous avons constaté une variabilité significative des performances en fonction des locuteurs. De plus, notre approche a été validée sur deux ensembles de données, produisant des tendances cohérentes et confirmant la généralisation des modèles.

Finalement, nous avons étudié la faisabilité de l'application des méthodologies développées dans un cadre clinique en détectant des états hypoglycémiques chez des patients diabétiques à partir de l'analyse de la parole. Dans ce but, nous avons utilisé des représentations neuronales provenant des réseaux d'estimation des schémas respiratoires, ainsi que d'autres représentations neuronales et caractéristiques acoustiques. Nous avons également examiné les performances des modèles d'estimation de la fréquence cardiaque dans ce contexte. Dans le cadre de cette recherche, nous avons compilé deux nouveaux ensembles de données avec des enregistrements simultanés de la parole et de signaux physiologiques, dont l'un a été collecté dans un environnement clinique. Ces ensembles de données ont été utilisés pour évaluer les performances des modèles développés.

Mots-clés : Traitement de la parole, estimation des schémas respiratoires, paramètres respiratoires, fréquence cardiaque, signaux physiologiques, apprentissage automatique, réseaux de neurones convolutifs, représentations neuronales, détection de COVID-19, hypoglycémie.

Contents

List of Figures

List of Tables

1 Introduction

Speech is the most common mode of communication among human beings. The human speech production system is capable of conveying a wide range of information – from the semantics of the spoken words to speaker dependent characteristics such as age (Bocklet et al., 2008), gender (Vergin et al., 1996), mental (Bedi et al., 2015), and emotional (El Ayadi et al., 2011) state. It can provide some information about the underlying physical condition (health) of a speaker as well (Orozco-Arroyave et al., 2015). This is due to the fact that speech production system is a complex combination of various systems which collaborate in perfect synchrony, and changes in any of them can result in variations in speech. It involves several physiological processes such as respiration, cardiac activity, and the control of muscle activity (Hardcastle et al., 2012; Minifie et al., 1973; Conrad et al., 1979). Additionally, other physiological factors, such as blood glucose levels, can influence vocal cord elasticity and, consequently, affect a person's speech production (Ulanovsky et al., 2014).

Traditionally, research in the domain of speech production has emphasized the activities of the voice source and the articulatory dynamics within vocal cavities. Recently, however, there has been a growing interest in integrating additional physiological signals into speech related studies. Most of these studies have focused on examining individual signals. For example, (Hammarsten et al., 2015) analysed breathing patterns in group communications. (Włodarczak et al., 2017) compared the breathing rate during speech and quite breathing and found that the breathing rate during speech is higher. They demonstrated that speech can also shape the breathing. (J. Smith et al., 2017) investigated the change in heart rate between when the person is speaking compared to when they are silent, and they showed that heart rate increases during speech specially when the speaker is frustrated. (Jati et al., 2018) predicted heart rate and respiratory sinus arrhythmia from speech during stressful conversations.

Despite the growing interest in this domain, to the best of our knowledge, prior to the beginning of this research, there has been limited effort to simultaneously model speech and physiological signals. In this thesis, we investigate the relationship between physiological signals and speech, with a focus on breathing pattern and heart rate. We integrate such physiological information into speech related applications. Alongside with the work presented in this

thesis, the Breathing sub-challenge of the Interspeech 2020 ComParE challenge (Schuller et al., 2020) was held, which focused on estimating breathing patterns from speech signals. Several methodologies were introduced during this challenge (Markitantov et al., 2020; Mendonça et al., 2020). Additionally, the Multimodal Sentiment Challenge (MuSe 2021 and 2022) provided simultaneous recordings of speech and physiological signals including respiration, heart rate, and skin conductance from participants in stressed dispositions. The challenge aimed to develop models for continuous emotion recognition using multimodal data (Stappen, Baird, et al., 2021; Christ et al., 2022). Most recently, (Ntalampiras, 2023) presented an ensemble method for estimating heart rate and breath rate from speech signals.

1.1 Motivation

Better understanding of the relation between speech and physiological signals is a valuable line of research. The motivation being:

1. As health care is moving from hospitals to domestic environments, new, automated health monitoring methods that are reliable, accessible, affordable, and user-friendly become more imperative. Physiological signals such as those related to respiration and cardiac function can indicate the underlying medical condition of a person. Such signals however might not be easily accessible as the traditional methods of recording such signals require specialized instruments and knowledge. Using alternative methods to acquire such physiological signals could be helpful when there is a need for long distance monitoring. Some medical conditions require continuous monitoring to prevent severe complications. For example, diabetic patients are in risk of hypoglycemia and need to monitor their blood glucose level, however the current methods are invasive, and expensive. Using speech to develop new methods for indirectly extracting physiological information is an attractive candidate, as speech recordings are readily accessible, cost-effective, and do not require specialized expertise. Moreover, some physiological signals such as breathing patterns are closely related to speech. Breathing provides the necessary airflow for speech production. Other physiological signals such as heart rate can be correlated to speech. For example, heart rate increases during speaking (J. Smith et al., 2017) and different emotional states can affect both speech and heart rate (A. P. James, 2015). Up to now there have been very few studies investigating the relationship between physiological signals and speech. Even fewer studies are taken place to estimate such physiological signals from speech.

2. Accurate estimation of breathing patterns and cardiac functions from speech can be applied in many areas. For instance, speech recordings from healthcare call centers could assist in monitoring a variety of physical and mental health conditions. Additionally, this approach could enhance current speech processing techniques for improved performance. For example, individuals with Parkinson's disease often experience dysarthria, a speech impairment affecting phonation, articulation, prosody, and intelligibility. Re-

search has predominantly focused on analysing articulation and prosody, as these aspects most visibly reflect dysarthria severity. However, phonation issues also play a vital role in assessing overall speech impairment in Parkinson's patients (Vásquez-Correa et al., 2021). Phonation, which involves producing vocal sounds by directing air from the lungs through the vocal cords, is inherently linked to breathing. Thus, estimated breathing patterns from speech could offer valuable insights into phonation impairments in Parkinson's patients, especially in cases where only speech data is available.

3. Better understanding of the relationships between speech and physiological signals can also make it possible to analyse speech and other physiological signals simultaneously. For example, using physiological signals such as breathing pattern and heart rate, when available, along with speech signals, can improve the automatic stress detection methods. An automatic stress detection method can be used to evaluate the stress level of personnel in critical conditions. Detecting stress in early stages could prevent risky health problems. Incorporating physiological signals with speech however requires a deeper understanding of the relationship between them.

The goal of this thesis is to investigate and understand the relationships between speech and physiological signals through machine learning, and to apply these insights to reinforce speech based applications.

1.2 Contribution

The contributions of this thesis are:

- **Estimating breathing patterns from speech signals.**
  Two deep learning based approaches have been used to develop models for estimating breathing patterns from speech signals: (a) spectral based approach in which spectral based features were first extracted from speech signals and used to train a RNN-LSTM model and a CNN model, and (b) end-to-end raw waveform based approach in which raw speech signals were used as input to a CNN model. Our focus has been on developing the raw waveform based models. We evaluated the models in single and cross database settings and investigated the limitations of the evaluation measures used in the study. This work has been done in collaboration with Philips Research and Center for Language Studies (CLS), Radboud University Nijmegen, The Netherlands. The scientific publications related to this work are as following:

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

Z. Mostaani, V. S. Nallanthighal, A. Härmä, H. Strik, and M. Magimai-Doss, "On the relationship between speech-based breathing signal prediction evaluation measures and breathing parameters estimation," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1345–1349.

Z. Mostaani, V. S. Nallanthighal, A. Härmä, H. Strik, and M. Magimai-Doss, "Estimating breathing pattern from raw speech waveform and short-term speech spectrum using neural networks," Idiap Research Report, Idiap-RR-12-2024, 2024.

- **Using breathing pattern estimation models in speech related applications and understanding their behavior.**
  We used the developed models in estimating breathing pattern in speech related applications such as COVID-19 detection, distinguishing between natural and synthetic speech, and emotion recognition. Additionally, we probed the models to understand the kind of information that is being modeled by them. The corresponding scientific publications include:

Z. Mostaani, R. Prasad, B. Vlasenko, and M. Magimai-Doss, "Modeling of pre-trained neural network embeddings learned from raw waveform for COVID-19 infection detection," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8482–8486.

Z. Mostaani and M. Magimai-Doss, "On breathing pattern information in synthetic speech," *Proc. Interspeech*, 2022, pp. 2768–2772.

S. Yadav, T. Purohit, Z. Mostaani, B. Vlasenko, and M. Magimai-Doss, "Comparing biosignal and acoustic feature representation for continuous emotion recognition," *Proc. of 3rd International Multimodal Sentiment Analysis Workshop and Challenge (MuSe' 22)*, 2022, pp. 37–45.

- **Estimating heart rate from speech signals.**
  We developed machine learning models to estimate cardiac metrics such as heart rate from speech signals. We used acoustic features and neural embeddings extracted from self-supervised learning models as input. We evaluated our models on two datasets. The corresponding scientific publications include:

G. Elbanna, Z. Mostaani, and M. Magimai-Doss, "Predicting heart activity from speech using data-driven and knowledge-based features," *Proc. Interspeech*, 2024, pp. 4758–4762.

- **Investigating a clinical application using the developed models.**
  We investigated a clinical application using the developed models both for estimating breathing patterns and heart rate from speech signals. We used those models along

with other select features to investigate the biomarkers in speech that are related to hypoglycemia (low blood sugar level) in diabetic patients and can be used to detect hypoglycemic state. This work has been done in collaboration with Swiss Center for Electronics and Microtechnology (CSEM) and Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism (UDEM), Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland.

- **Collecting synchronized speech and physiological data.**
  Studying the relationship between speech and physiological signals required speech and physiological data that has been collected in a synchronized manner. As part of this thesis, we have collected two datasets that includes speech and physiological signals. In both datasets, several speech tasks were performed by the participants. The first dataset includes only healthy subjects which perform the tasks in normal conditions. The second dataset, collected in a clinical setting, includes data from diabetic patients which perform the tasks while their sugar level is in normal and hypoglycemic state. The device required to collect the physiological signals was developed by CSEM SA.

1.3 Thesis Outline

The rest of the thesis is organized as follows:

In Chapter 2, we provide an overview of the machine learning methods used in this thesis, including classical approaches and deep learning based techniques. The type of features utilized are also introduced in this chapter.

In Chapter 3, we present the work on estimating breathing patterns from speech signals. The models developed for breathing pattern estimation are used in Chapter 4 for three different applications, namely, COVID-19 detection, distinguishing between natural and synthetic speech, and emotion recognition. We further probe the models to understand the type of information that is being modeled by them in Chapter 5.

In Chapter 6, we present methods to estimate heart rate from speech signals. The models that have been developed for extracting physiological information from speech signals along with other select features are used in a clinical setting in Chapter 7. We investigate the feasibility of detecting hypoglycemic state in diabetic patients based on speech signals. Finally, in Chapter 8, we conclude the thesis and discuss future directions.

2 Background

The goal of this thesis is to extract physiological information from speech signals using machine learning techniques and using physiological information alongside speech features in various applications. Towards this goal, we employed a range of machine learning techniques, feature extraction methods, and end-to-end learning approaches in addressing both regression and classification problems. In this chapter, we introduce key concepts in machine learning that will be utilized throughout this thesis.

Extracting physiological information from speech signals can be categorized as paralinguistic analysis of speech. A high level overview of a typical machine learning pipeline used in this domain is illustrated in Figure 2.1. This standard approach is modular: initially, features are extracted from the speech signal, and these features are subsequently fed into a machine learning model to perform a designated task. The target task could be either a classification or a regression problem. In this thesis, we implemented such a pipeline to predict heart rate from speech signals as a regression problem. In another application, we estimated the arousal and valence of the speaker from speech signals, also framed as a regression problem. Additionally, we utilized the pipeline in classification tasks, including COVID-19 detection, distinguishing between natural and synthetic speech, and the detection of hypoglycemic state from speech signals.

Another approach in speech processing that has gained significant attention in recent years

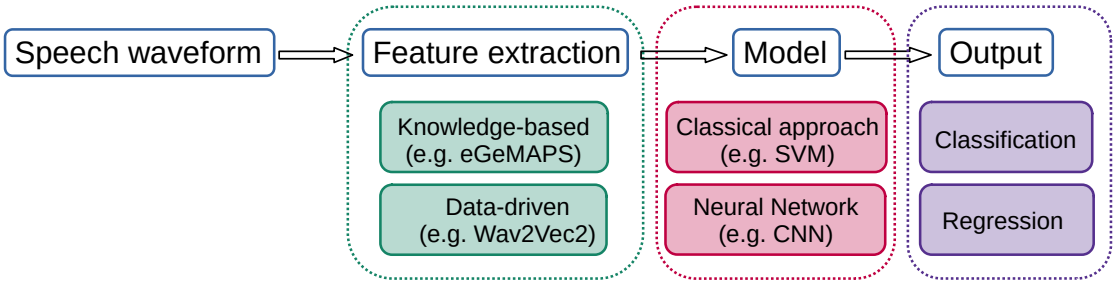


Figure 2.1 – Modular approach for a machine learning pipeline

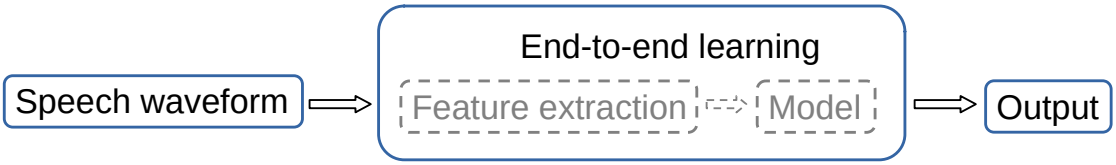


Figure 2.2 – End-to-end learning approach for a machine learning pipeline

is end-to-end learning. In this paradigm, unlike the modular approach, features are not explicitly extracted from the speech signal. Instead, the raw speech signal is directly input into a neural network, which learns to extract features and train a model within a single pipeline (See Figure 2.2). In this thesis, we applied end-to-end learning techniques to extract breathing patterns from speech signals.

In the remainder of this chapter, we will provide further details about both the modular approach and the end-to-end learning methods. We will also discuss various evaluation metrics that are commonly used for regression and classification tasks.

2.1 Modular Approach

In this section, we will discuss the various components of the machine learning pipeline in a modular approach. The pipeline consists of feature extraction methods, machine learning models, and target tasks.

2.1.1 Feature extraction

Speech signals are quasi-stationary, meaning they can be considered stationary over short periods of time. Typically, features are extracted from short segments of speech signals (around 20-30 ms), referred to as frames. These frame-level representations serve as input to machine learning models. In paralinguistic analysis, it is common to aggregate these frame-level representations over longer durations to obtain a fixed-length utterance-level representation for model input (see Figure 2.3). A common aggregation method involves calculating statistics over the frames throughout the duration of the utterance (functionals). For instance, we used first and second order statistics (mean and standard deviation) as aggregation method in various applications. Another technique employed in this thesis is computing bag-of-audio-words (BoAW) (Pokorny et al., 2015). In this method, first the frame-level representations are clustered into a fixed number of clusters (codebook), with the centroid of each cluster serving as a codeword. Each utterance is then represented by a histogram of these codewords, indicating the frequency of different patterns in the signal.

Feature extraction methods can be classified primarily into two categories: knowledge-based features and data-driven features, both of which are utilized throughout this thesis.

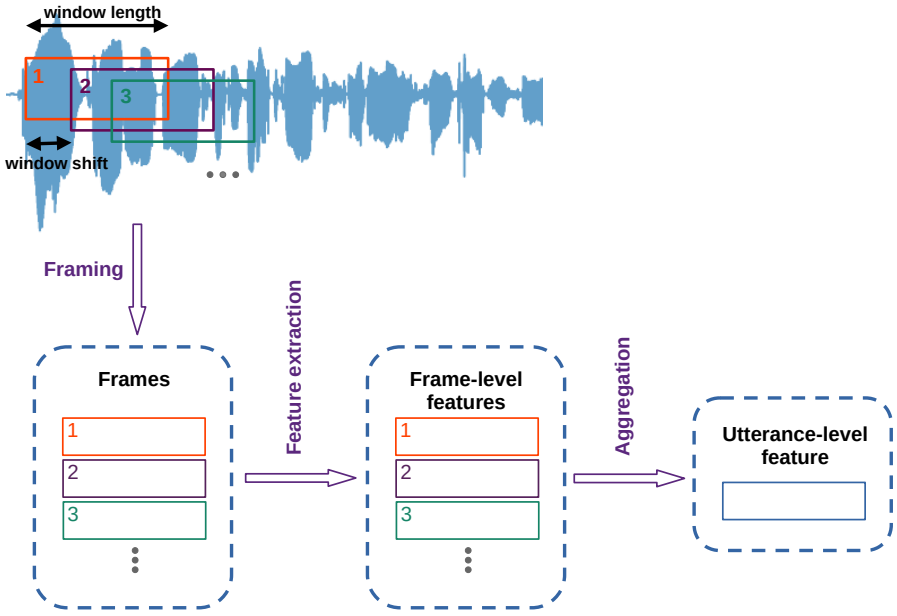


Figure 2.3 – Frame-level and utterance-level representations.

2.1.1.1 Knowledge-based features

Knowledge-based features are derived from expert understanding in domains such as voice acoustics, phonetics, and paralinguistics. Below, we outline four feature sets utilized in this thesis.

- **Computational Paralinguistics Challenge (ComParE):** The ComParE feature set comprises a rich set of features derived from low-level descriptors (LLDs) such as pitch, loudness, jitter, shimmer, and Mel-Frequency Cepstral Coefficients (MFCCs). These features capture various paralinguistic cues from speech (Schuller, Steidl, Batliner, et al., 2013). ComParE LLDs include 65 descriptors which are extracted on a frame-by-frame basis. Additionally, their approximate first and second temporal derivatives, referred to as delta and delta-delta, can be computed in a frame-wise manner. A set of functionals (e.g., mean, standard deviation, skewness) are applied to the frame-level descriptors to obtain utterance-level representations (see Figure 2.3) which yield a 6373-dimensional feature vector (Eyben et al., 2013). ComParE features have demonstrated effectiveness in numerous paralinguistic tasks, including emotion recognition (Schuller, Steidl, Batliner, et al., 2013), the detection of neurological disorders like Parkinson's disease (Schuller et al., 2015), and emerging applications such as COVID-19 detection from speech (Schuller et al., 2021).
- **Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS):** Unlike ComParE, which is a large feature set, eGeMAPS is a minimalistic collection of audio parameters carefully selected for their potential to detect physiological changes in voice production, their previous application in studies, and their theoretical significance (Eyben et al.,

2016). The eGeMAPS feature set includes frame-level LLDs such as pitch, jitter, and shimmer, and MFCCs which is then aggregated over the duration of an utterance using various functionals, resulting in an 88-dimensional feature vector. The effectiveness of eGeMAPS features has been established in tasks such as depression detection (Mamidisetti et al., 2023), stress detection (Baird, Triantafyllopoulos, et al., 2021), as well as evaluations of vocal disorders (Barche et al., 2020) and cognitive impairments (Haider et al., 2020).

- **Spectogram:** A spectrogram represents the frequency spectrum of a signal over time. It is widely used in audio and speech analysis to visualize changes in energy among various frequency components over time. Spectrograms are obtained by applying a short-time Fourier transform (STFT) to speech frames. Log Mel spectrograms transform the frequency axis according to the Mel scale to better align more closely with human auditory perception. The magnitudes of spectral components are logarithmically transformed, capturing perceptually relevant features of speech. Spectograms and Log Mel spectograms have been used in many speech related tasks, such as, speaker recognition (Z. Liu et al., 2018), emotion recognition (Satt et al., 2017) and pathological speech processing (Gallardo-Antolín et al., 2021; Vásquez-Correa et al., 2017).

- **Long-term average spectrum (LTAS):** LTAS averages the spectral information of an audio signal over an extended duration, typically across entire speech or audio recordings. In contrast to the time-varying information provided by spectrograms, LTAS provides a view of the overall energy distribution across the frequency spectrum for the entire recording. LTAS techniques have been applied in clinical settings and utilized as quality measures in pathological speech analysis (Löfqvist, 1986; Master et al., 2006; L. K. Smith et al., 2014; Halpern et al., 2021). It has also been used in speaker verification systems to detect presentation attacks (Muckenhirn et al., 2017).

2.1.1.2 Data-driven features

Data-driven feature extraction methods depend on learning representations directly from data, rather than totally relying on preprocessing techniques to extract features based on predefined assumptions. In this approach, pre-trained neural networks are employed to derive meaningful representations from raw data. The intermediate representations learned by neural networks, referred to as *embeddings*, are vectors generated from the intermediate layers of the network. These embeddings represent abstract features of the input data. Pre-trained models can be employed without fine-tuning, allowing the extracted embeddings to serve as input for subsequent machine learning models tackling downstream tasks. Similar to knowledge-based features, embeddings can be extracted from speech frames before being used as input to machine learning models, or aggregated to generate a fixed-length representation for each utterance.

Embeddings may be derived from neural networks trained in a supervised manner for aux-

iliary tasks. In this case, the neural networks are originally trained to solve a specific task, using labeled data. The embeddings learned by these networks, which capture task specific information, are then used as input features for other tasks. For instance, embeddings obtained from networks trained for phoneme classification have proven effective for emotion recognition(Purohit, Vlasenko, et al., 2023). Another approach in training neural networks which has gained attention is self-supervised learning. In this method, unlike the supervised learning which requires labeled data, the neural networks learn meaningful representations using unlabeled data. Instead, they learn labels from the data itself, to solve a task. Recently, self-supervised learning models trained on large datasets have gained attention (Mohamed et al., 2022), as they capture generalized representations that are applicable in various tasks, such as depression detection (W. Wu et al., 2023), stress detection (Elbanna, Biryukov, et al., 2022), and dysarthric speech classification (Javanmardi et al., 2023). In this thesis, we employed both supervised and self-supervised learning models without fine-tuning to extract embeddings from speech signals.

2.1.2 Machine learning models

Machine learning models can be trained using a variety of algorithms. The advent of deep learning has resulted in widespread application of neural networks. We categorize training approaches into two primary types: classical approach and artificial neural networks.

2.1.2.1 Classical approach

Classical approaches refer to traditional machine learning techniques that do not involve deep learning or neural networks. These methods rely on well-established algorithms and statistical principles. In contrast to deep learning, which typically requires large datasets and complex neural network architectures, classical methods use simpler models that can effectively function with smaller datasets. These models have been employed in paralinguistic studies due to data scarcity in this domain. Below are examples of classical models used in this thesis.

- **Linear regression:** Linear regression is a fundamental model that describes the relationship between input and output as a linear function (G. James et al., 2023). The learning algorithm minimizes the error between predicted and actual outputs by finding the best hyperplane that fits the data. One commonly used error function is the least squared error, which minimizes the summation of squared errors for each data point. Ridge regression incorporates L2 regularization into the linear regression model to prevent overfitting on the training data.
- **Logistic regression:** Used for classification tasks, logistic regression finds the optimal hyperplane that separates data into classes (G. James et al., 2023). In binary classification, the probability of an input belonging to a certain class is modeled using a sigmoid

function.

- **Support vector machines (SVM):** SVM models find the best hyperplane that maximizes the margin between the hyperplane and the closest data points (support vectors) to it (Vapnik, 1963; Cortes et al., 1995). Originally SVMs were designed for linearly separable data. Higherowever, with the introduction of kernel methods they can handle non-linearly separable data (Boser et al., 1992). This thesis employs SVMs with linear, polynomial, and Radial Basis Function (RBF) kernels.

- **Decision trees:** Decision trees are simple models that make decisions based on conditional statements, represented in a tree structure. Decision nodes outline the conditions, and branches represent possible outcomes. Leaf nodes signify the final decisions. While decision trees tend to overfit training data, ensemble learning techniques, such as Random Forest (RF), mitigate this issue. In RF, multiple small decision trees (weak learners) are trained using random subsets of data samples and features (Ho, 1995; Breiman, 2001), with final decisions made by aggregating the outputs of all trees. Other ensemble methods employed in this thesis include Gradient Boosting (GB) (Hastie et al., 2001) and AdaBoost (AB) (Freund et al., 1996).

2.1.2.2 Artificial neural networks

Artificial Neural Networks (ANNs) were first introduced in the 1940s but have gained popularity in recent decades due to advancements in computational power and the availability of large datasets. ANNs are widely utilized in various applications, including image recognition, speech recognition, and natural language processing (Goodfellow et al., 2016).

Different architectures for artificial neural networks exist, typically composed of layers of interconnected neurons with associated weights. These networks are commonly trained using the backpropagation algorithm (Rumelhart et al., 1986). This process involves defining an error between the actual and predicted outputs and propagating this error backward through the network. Weights associated with the connections are updated to minimize this error as it moves from the output layer back to the input layer. The error measure (also referred to as cost function) is defined based on the task at hand. For example, in regression tasks, the Mean Squared Error (MSE) is commonly used whereas in classification tasks, a common error measure is the Cross-Entropy. Stochastic Gradient Descent (SGD) (Bottou, 2010) and the Adam optimizer (Kingma et al., 2015) are examples of optimization algorithms used in conjunction with backpropagation to minimize errors. While this thesis does not focus on the optimization algorithms employed for training neural networks, we utilize various neural network architectures to extract physiological information from speech signals. The following structures have been applied in this thesis:

- **Multi-layer perceptron (MLP):** An MLP is a feedforward neural network consisting of an input layer, multiple hidden layers, and an output layer. Each neuron in a layer connects

to all neurons in the subsequent layer, with weights and biases associated with these connections.

- **Convolutional neural network (CNN):** CNNs are specialized neural networks designed to capture localized information from grid-like data, such as images and time-series (LeCun et al., 1995). They consist of several convolutional layers, where convolution filters are applied, followed by pooling layers that reduce dimensionality while retaining critical information. CNNs act as feature extractors and are typically followed by fully connected layers for task execution.

- **Recurrent neural network (RNN):** RNNs are designed to capture temporal patterns in data using feedback loops in their architecture (Rumelhart et al., 1986). Long short-term memory (LSTM) networks, a specific type of RNN, address the vanishing gradient problem, enabling them to capture long-term dependencies in the data (Hochreiter et al., 1997).

2.2 End-to-End Acoustic Modeling

In Section 2.1.1, we discussed feature extraction methods for deriving representations from speech signals. Both knowledge-based and data-driven feature extraction methods require certain assumptions about the data to generate meaningful features. However, the advent of deep learning has prompted a transition towards *end-to-end learning* in speech signal analysis. In this framework, neural networks directly process raw speech waveforms as input, without requiring assumptions regarding beneficial information for specific tasks. End-to-end models are trained to automatically learn both feature representations and the task (classification or regression) through a unified pipeline.

We based our end-to-end models on a CNN architecture, which has demonstrated considerable success in numerous speech related tasks. The architecture has been originally proposed for speech recognition (Palaz et al., 2013) and has since been explored in various speech processing tasks such as speaker recognition (Muckenhirn et al., 2018), gender recognition (Kabil et al., 2018), depression detection (Dubagunta et al., 2019) and more recently for emotion recognition (Purohit, Yadav, et al., 2023). The CNN architecture consists of several convolution and max-pooling layers followed by a fully connected layer (MLP), and finally, an output layer. The convolution layers function as feature extractors, while the fully connected layer operates as a classifier. Each convolution layer includes a set of convolutional kernels that are applied either to the input speech signal or to the output of the previous layer. The max-pooling layers reduces the dimensionality of the data by retaining the maximum value from a neighboring set of values. An activation function is usually applied to introduce non-linearity into the system. Figure 2.4 depicts a convolution layer with C sliding kernels applied to input. The kernel width is determined by kw, and the stride is denoted by dk. The outputs are forwarded to a max-pooling layer characterized by width mw and stride dm. The kernels in the convolution layers are learned during the training process. The kernels of the first convolution layer can be

considered as filters (Palaz et al., 2019). The cumulative frequency response of these filters have been investigated to understand how the source and system related information in the speech signal is modeled when using this approach (Palaz et al., 2019; Muckenhirn et al., 2018; Muckenhirn et al., 2019).

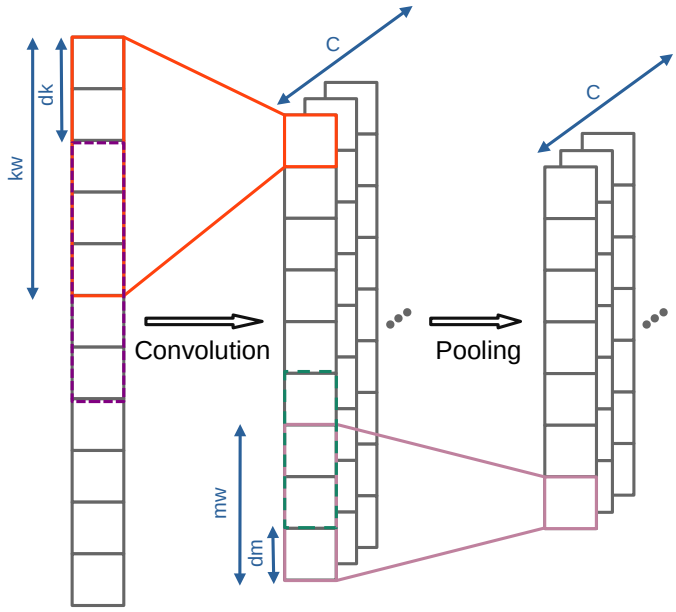


Figure 2.4 – A one dimensional convolution and pooling.

In this thesis, unlike the previously mentioned works which used this architecture for a classification task, the end-to-end modeling is applied in a regression framework to extract breathing patterns from speech signals. Additionally, we analyse the cumulative frequency response of the learned filters for the first convolution layer to gain insight into the type of information modeled by the network for the breathing pattern estimation task.

2.3 Evaluation Measures

Regardless of the method used for training a machine learning model, it is essential to evaluate the model's performance. The evaluation metrics used depend on the specific task at hand, whether it is a regression or classification problem. In this section, we discuss the evaluation metrics used in this thesis.

2.3.1 Classification

Classification tasks aim to predict discrete class labels for given input data. It is also possible to predict the probabilities for each class, which can subsequently be converted into class labels using a specified threshold. The classifier can be trained for binary classification (distinguishing between two classes) or multi-class classification (distinguishing among more

than two classes). In a detection problem, which is a specific type of binary classification, the focus is on identifying the presence or absence of a particular class. Essentially, it involves distinguishing between the target class (of interest) and all other instances. In this thesis, we focus on detection problems. The following evaluation metrics are commonly used in detection tasks:

- **Area under the receiver operating characteristic curve (AUC):** AUC is a widely used metric that quantifies the model's ability to distinguish between positive and negative classes. It measures the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various threshold settings. AUC values range from 0 to 1, with higher values indicating better model performance.
- **Equal error rate (EER):** EER is the point on the ROC curve where the false positive rate equals the false negative rate. It provides a balance between the two error rates, making it a useful metric for imbalanced datasets. The lower the EER, the better the model performance.
- **Sensitivity:** Sensitivity, also known as the true positive rate or recall, measures the proportion of actual positive samples correctly identified by the model. It is calculated as the ratio of true positives to the sum of true positives and false negatives.
- **Specificity:** Specificity, also known as the true negative rate, measures the proportion of actual negative samples correctly identified by the model. It is calculated as the ratio of true negatives to the sum of true negatives and false positives.
- **Precision:** Precision or positive predictive value, measures the proportion of predicted positives that are actual positives. It is calculated as the ratio of true positives to the sum of true positives and false positives.
- **F1-score:** The F1-score is the harmonic mean of precision and recall. The F1-score ranges from 0 to 1, with higher values indicating better model performance.

2.3.2 Regression

Regression tasks involve predicting continuous values for input data. In this thesis, we focus on regression problems to estimate physiological signals from speech features. The following evaluation metrics are commonly used in regression tasks:

- **Mean squared error (MSE):** MSE measures the average squared difference between the observed values and the predicted values. It provides a measure of the model's accuracy in predicting continuous values. The lower the MSE, the better the model performance. MSE is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2.1}$$

where:

- Y_i are the observed values,
- $\hat{Y}_i$ are the predicted values from the model,
- n is the number of data points.

- **Mean absolute error (MAE):** MAE measures the average absolute difference between the observed values and the predicted values. It provides a robust measure of the model's performance, as it is less sensitive to outliers compared to MSE. The lower the MAE, the better the model performance. MAE is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \tag{2.2}$$

where:

- Y_i are the observed values,
- $\hat{Y}_i$ are the predicted values from the model,
- n is the number of data points.

- **Pearson's correlation coefficient (PCC):** The Pearson's correlation coefficient, denoted by r, quantifies the strength of the linear relationship between two variables. When applied to time series data it measures how well the two signals correlate at the same time points. The coefficient ranges between -1 and $+1$, where $+1$ indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. r is given by:

$$r = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}} \tag{2.3}$$

where:

- Y_i are the observed values,
- $\hat{Y}_i$ are the predicted values from the model,
- $\bar{Y}$ and $\bar{\hat{Y}}$ represent the means of Y and $\hat{Y}$ respectively,
- n is the number of data points.

This coefficient captures how well the changes in one variable (the predictions) are associated with the changes in the other variable (the observations).

- **Concordance correlation coefficient (CCC):** The concordance correlation coefficient, denoted by ρ_c, assesses the degree to which the predicted values ($\hat{Y}$) align with the observed values (Y). It ranges from -1 to $+1$, where values approaching $+1$ indicate higher concordance between predictions and observed values. ρ_c is given by:

$$\rho_c = \frac{2\,\mathrm{cov}(Y, \hat{Y})}{\sigma_Y^2 + \sigma_{\hat{Y}}^2 + (\bar{Y} - \bar{\hat{Y}})^2} \tag{2.4}$$

  where:

  - Y are the observed values,
  - $\hat{Y}$ are the predicted values from the model,
  - $\mathrm{cov}(Y, \hat{Y})$ is the covariance between Y and $\hat{Y}$,
  - σ_Y^2 and $\sigma_{\hat{Y}}^2$ are the variances of Y and $\hat{Y}$ respectively,
  - $\bar{Y}$ and $\bar{\hat{Y}}$ are the means of Y and $\hat{Y}$ respectively.

- **Coefficient of determination (R^2):** The coefficient of determination, R^2, is a measure that how well a model predicts an outcome. It is calculated as the proportion of total variation of model outcomes. The R^2 value ranges from 0 to 1, with higher values indicating a better fit of the model to the data. In case of non-linear models, the R^2 value can be negative. R^2 is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{2.5}$$

  where:

  - Y_i are the observed values,
  - $\hat{Y}_i$ are the predicted values from the model,
  - $\bar{Y}$ is the mean of the observed values,
  - n is the number of data points.

  The numerator represents the residual sum of squares (the deviation between the observed values and the predicted values), and the denominator represents the total sum of squares (the deviation of the observed values from their mean). Higher R^2 values indicate that the model explains a larger portion of the variation in the observed data.

2.4 Conclusion

In this chapter, we provided an overview of key concepts in machine learning relevant to this thesis. We introduced the two primary approaches in machine learning, modular approach and end-to-end learning. We discussed the feature extraction methods used in the modular

approach and gave an overview of the classical and artificial neural network models employed in this thesis. The end-to-end learning approach employed in this thesis was introduced. We also discussed the evaluation metrics used for regression and classification tasks. In the following chapters we will apply these concepts to tackle specific challenges related to the extraction of physiological information from speech signals.

3 Breathing Pattern Estimation

3.1 Introduction

The COVID-19 pandemic demonstrated the necessity of remote digital health assessment tools like telehealth monitoring services for sustainable health care, particularly for elderly and vulnerable populations. Speech is a good indicator of the pathological condition of a person (Dibazar et al., 2002). Speech production is a complex process involving perfect synchrony among various systems like muscular, respiratory, cognitive, and autonomic nervous systems. Any lapse in any of these systems could significantly affect one's speech. The use of speech analytics has been gaining attention within the clinical and health care domains in recent years. This follows to the success of deep learning techniques in various speech technology applications (Miotto et al., 2018) and speech pathology (Cummins et al., 2018; Koolagudi et al., 2018), and has been replacing the conventional knowledge-based machine learning techniques (Cummins et al., 2017; Teixeira et al., 2013).

Speech and respiration are closely related. Speech is produced by organs evolved for the respiratory function of the body (MacLarnon et al., 1999). Breathing is a primary mechanism of speech generation. The special mechanism of using the respiratory system to produce the airflow necessary for phonation is termed as *speech breathing* (Von Euler, 1982). Speech breathing is implicated in many aspects of speech production, such as voice quality (Slifka, 2006), voice onset time (Hoit et al., 1993), and loudness (Huber et al., 2005). Vocalization mostly takes place during exhaling while inhaling is done in quick pauses in between utterances. We subconsciously exercise continuous breathing planning during the speech. Breathing planning is evident as we take in more air for a long continuous utterance (Włodarczak et al., 2017). Respiratory diseases such as chronic obstructive pulmonary disease (COPD), asthma, and respiratory infection are common in elderly populations. These conditions significantly influence the breathing capacity and thus influence breathing planning resulting in frequent breaks in an utterance in need of air. We can hear when a person has breathing difficulties, but the automatic detection is a complex task because the breathing planning is based on linguistic and prosodic factors (Włodarczak et al., 2015).

The current research is related to the development of acoustic sensing technology for telehealth services related to respiratory conditions in particular. Breathing monitoring from telehealth customers' speech conversations over multiple calls would give us the historical data of breathing parameters and help us compare and understand a person's pathological condition, decline, or improvement over time and early detection of a condition.

In this chapter, we present a comprehensive research study on deep learning architectures for estimating breathing patterns and breathing parameters directly from speech and address the following challenges to establish the respiratory analysis of speech as a practical or clinical application.

1. Firstly, we establish the possibility of estimating breathing signal from the speech signal in two ways: the spectral analysis and raw waveform analysis. We then propose a fusion method to enhance our system's performance. This study has been done in collaboration with Philips Research, Eindhoven, The Netherlands. While my focus has been on developing the raw waveform based models, the spectral features based models have been developed in parallel by my collaborator.

2. To make the study more robust and reliable for practical or clinical purposes, we study two significant protocols: read speech and conversational (or spontaneous) speech. We perform the study on individual databases. We further perform a cross database analysis to assess how our systems generalize. This comprehensive study helps in applying this technology independent of databases.

3. For evaluating the estimated breathing signal from the neural network models, we use standard evaluation metrics for the regression problem: mean squared error and correlation with reference to actual breathing signal measured through respiratory inductive belts as discussed in Section 3.4.4. However, the practical utility would be justified by comparing the breathing parameters like breathing rate, tidal volume equivalent, and breath event sensitivity.

4. We investigate whether the evaluation measures used to assess breathing signal prediction models are sufficient indicators for reliable breathing parameter estimation.

5. In some cases, breathing sounds are also audible in a speech recording. In such cases, it is easier to detect breath events. Ruinskiy and Lavner proposed an effective breath event detection algorithm based on template matching for automatic detection and exact demarcation of breath sounds during speech (Ruinskiy et al., 2007). However, in the experiments reported in this chapter, breathing sounds are not recorded or used in the analysis, which is ensured by recording speech at a distance from the speaker's mouth during data collection, as described in Section 3.4. The respiratory sensing from the speech is based on the composition of speech utterance, i.e., linguistic content and prosodic factors independent of breath sounds.

The remainder of the chapter is organized as follows. We provide a background on the relation between speech and respiration and breathing parameter estimation from breathing signal in Section 3.2. In Section 3.3, we present the different approaches investigated for estimating breathing signal from speech, which includes neural networks based on spectral features input and raw speech input with various loss functions. Following this, in Section 3.4, we detail the experimental setup and databases used. In Section 3.5, we present results of different experimental studies. In Section 3.6, we present an analysis of the studied approaches. Finally, in Section 3.7, we conclude the chapter. The material presented in this chapter was originally published in modified form in (Nallanthighal, Mostaani, et al., 2021; Mostaani et al., 2021; Mostaani et al., 2024).

3.2 Background

Very few studies are focused on the relationship of speech and respiration and the effect of speech on breathing pattern in recent years. Breathing patterns provide medical doctors and speech therapists vital information about an individual's respiratory and speech planning (Székely et al., 2020), as well as cognitive and neurological health (Mitchell et al., 1996; Heck et al., 2017). J.D. Hoit and T.J. Hixon conducted early studies and explored different aspects of speech breathing like age (Hoit et al., 1987), body type (Hoit et al., 1986), gender (Hoit et al., 1989), and pathological conditions with neuromotor disorders to evaluate respiratory control in individuals (Solomon et al., 1993). (Winkworth et al., 1994) investigated the associations between linguistic factors and lung volumes in read speech and concluded that speech breathing is subject to a number of linguistic and prosodic influences. The amount of air breathed in and the volume of air in the lungs have been shown to be strongly influenced by the length and loudness of the intended utterance, whereas the expiratory duration is largely determined by the linguistic intent and this expiration can be modeled as a composition of phonemes with varying exhaustion flows for vowel and consonant phonemes (Klatt et al., 1968). (Hammarsten et al., 2015) investigated the inhalation duration and speech onset delay in different settings and reported that both of them are longer when speakers start to speak compared to when they are in the middle of a conversation. In other works, the breathing pattern for read speech has been compared to spontaneous speech. They reported that a high percentage of the sentences in read speech is produced during one breath while the inhalations were short and frequent during spontaneous speech (Y.-T. Wang et al., 2010; Henderson et al., 1965; Winkworth et al., 1994). The later could be due to the cognitive load during spontaneous speech (Mitchell et al., 1996).

When utilizing breathing for the purpose of the speech, the rate and volume of inhalation and the rate of exhalation during the utterance seem to be governed largely by the speech controlling system and its requirements with respect to phrasing, loudness, and articulation. Speech breathing demands more effort than normal quiet breathing. Quiet breathing encompasses relatively equal phases of inhalation and exhalation in terms of duration, amplitude, and velocity, whereas speech breathing is characterized by short inhalations to minimize inter-

ruptions to the speech flow and long exhalations due to higher resistance in the upper airway that prevent air from quickly flowing out (Hixon, 1987). Thus during speech, the breathing rate is approximately halved compared to quiet breathing (Włodarczak et al., 2017; Nallanthighal et al., 2019). (Włodarczak et al., 2017) proposed that the relationship between speech and breathing is not one way and breathing can also shape the speech. Thus, speech and breathing are closely related and gives an intuition for our hypothesis that speech breathing pattern can be sensed from the linguistic content and prosodic factors of the speech. This hypothesis has also been the basis for Breathing sub-challenge of Interspeech 2020 ComParE challenge (Schuller et al., 2020).

Both the rib cage and the abdomen can be used to modulate alveolar pressure and airflow during speech. Some speakers exhibit the more vigorous use of rib cage over abdominal contributions, and some speakers show a relatively equal contribution from both the rib cage and the abdomen (Hixon et al., 1976). The chest wall has been treated as a two-part kinematic system comprised of the rib cage and diaphragm-abdomen in parallel with only one degree of freedom each (Konno et al., 1967), and wherein the volume displaced by each part is linearly related to the motions of points within it (Hixon et al., 1976). When a known air volume is inhaled and measured with a spirometer, a volume-motion relationship can be established as the sum of the abdominal and rib cage displacements (Konno et al., 1967).

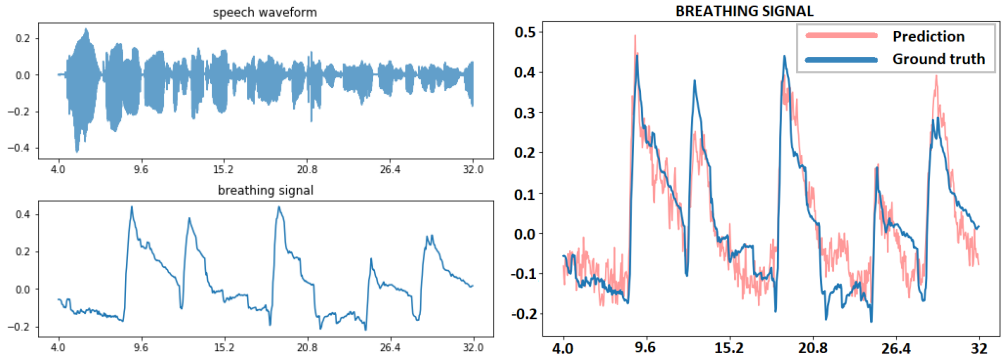


Figure 3.1 – (a) The speech waveform and corresponding breathing signal and (b) the predicted and ground truth for breathing signal.

Breathing signals are analysed to get breathing rate and tidal volume equivalent, which are the essential respiratory parameters to detect a person's pathological condition. These parameters are compared for the actual and estimated sensor data to determine the accuracy of estimation.

1. *Breath event* is the event of inhalation, which marks the beginning of the breathing cycle. During speech, we observe a sharp peak during inhalation and gradual decline during exhalation, as shown in Figure 3.1. This quasi-periodic pattern repeats over the course of a speech utterance. These peaks are determined using the automatic multiscale peak detection (AMPD) algorithm (Scholkmann et al., 2012), which is particularly relevant for detecting peaks in noisy periodic and quasi-periodic signals. Breath events of both

actual and predicted breathing signals are compared to evaluate the overlap of breath events, which ensures better prediction. This overlap is evaluated by the sensitivity of breath events.

2. *Breathing rate* is the average number of breath events per minute (Fuchs et al., 2015) and is computed by using AMPD algorithm (Scholkmann et al., 2012).

3. *Speech tidal volume* is a measure of the amount of air a person inhales during a normal breath for speech. It gives information about the lung capacity of a person (Konno et al., 1967). We normalized the average area under the curve per breath and used it as a tidal volume equivalent. This normalized tidal volume equivalent is used to compare actual and estimated breathing signals. We consistently use the term "tidal volume" to describe the above-mentioned speech tidal volume equivalent in this chapter.

3.3 Approaches

In the previous section, we observed that there exists a relationship between speech and respiration. This suggests that breathing signal could be estimated from the speech signal. Estimating breathing signals from the speech signal can be formulated as a regression problem. Such a problem could be formulated as a feature vector-to-feature vector regression problem (Qi et al., 2020; Qi et al., 2019) or feature vector-to-signal regression problem (Xu et al., 2014) or signal-to-signal regression problem (Fu et al., 2017; Rethage et al., 2018; Sebastian et al., 2020). The feature vector-to-feature vector regression formulation presumes that there exists a mathematical model for the signals based on the regressed features. In the case of speech signal, such a model is based on source-system decomposition through short-term analysis (Makhoul, 1975; Oppenheim et al., 2004; Ou et al., 2012). However, defining such a feature-based model for breathing signal is not trivial. Feature vector-to-signal regression and signal-to-signal regression problems have largely focused on problems where the input and output signals are of the same kind. For instance, in speech enhancement the input is corrupted speech signal and the output is clean speech signal (Fu et al., 2017; Rethage et al., 2018). In the case of breathing signal estimation from speech signal, we are dealing with two different kinds of signals. Furthermore, although there exists a relationship between speech and respiration, this relationship has not been fully characterized. In the sense that we do not know exactly which properties of speech signal characterize breathing signal. So, in the present work, we approached breathing signal estimation in two different ways using deep learning:

1. Apply short-term spectral processing on the input speech signal and then map the resulting representations to breathing signal. We refer to it as spectral features based approach. Here again we consider two sub-approaches. First, where the envelope of short-term spectrum is extracted and modeled with temporal context to estimate breathing signal. Second, where no such prior knowledge is applied and the short-term

Fourier magnitude spectrum with temporal context is modeled to estimate breathing signal. This can be regarded as a feature vector-to-signal regression formulation.

2. Learn to predict breathing signal directly from raw waveforms. We refer to it as raw waveform based approach. This can be regarded as a signal-to-signal regression formulation.

The motivation to use deep learning comes from the fact that we have less prior knowledge about the problem. Deep learning is capable of tackling lack of prior knowledge (Goodfellow et al., 2016).

3.3.1 Spectral features based approach

Spectral features are based on a time-frequency decomposition of the speech signal (see Section 2.1.1.1). We used the spectrogram and log Mel spectrogram to represent the spectral features of the speech signals as inputs to neural networks.

1. **Spectrogram:** We used the spectrogram generated by a short-time Fourier transform (STFT) with short frame size of 25ms and stride of 10ms (Sejdić et al., 2009). The Hamming window is applied to each frame and STFT is computed to get the power spectrum.

2. **Log Mel spectrogram:** Mel filter banks (n=40) were applied to the power spectrum to get the Mel spectrum. Mel filter banks use Mel frequency scaling, which is a perceptual scale to replicate human ear perception of sound (Stevens et al., 1937). It corresponds to better resolution at low frequencies and less at high frequencies.

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \tag{3.1}$$

$$f = 700(10^{\frac{m}{2595}} - 1) \tag{3.2}$$

where f is frequency in Hertz and m is Mel scale

Spectrogram and log Mel spectrogram of a speech signal of a fixed time window is mapped with respiratory sensor value at the endpoint of the time window with a stride of 10ms between windows for Philips database and a stride of 40ms between windows for UCL-SBM database to train the neural network models as shown in Figure 3.2. These models will estimate the respiratory sensor values of a speech signal in real-time to get the breathing pattern.

Using spectral features as an input representation of speech signal, we implement convolutional neural network (CNN) and long short-term memory recurrent neural network (LSTM-RNN) models (see Section 2.1.2.2) using the PyTorch software framework (Paszke et al., 2019).

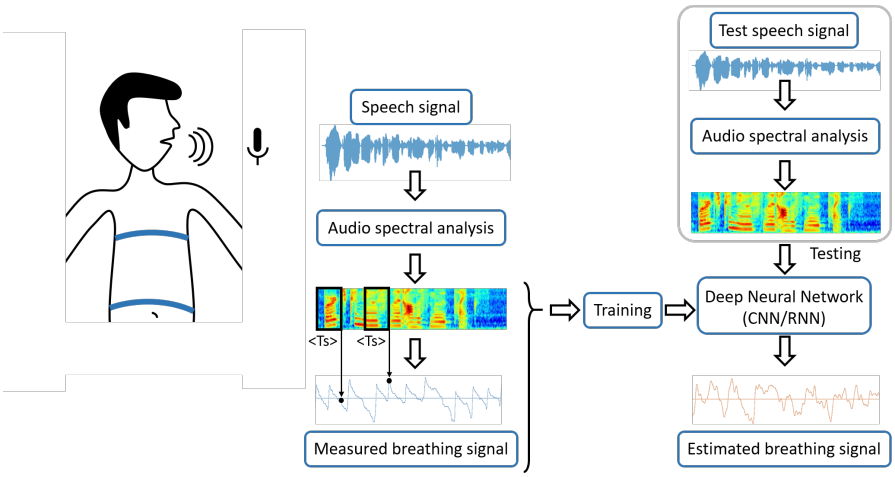


Figure 3.2 – Schematic diagram for estimating respiratory signal using deep neural network model based on spectral features.

CNN Model Input: log Mel spectrogram or spectrogram *m*: frames in time window *n* : Mel filter banks	**RNN Model** Input: log Mel spectrogram or spectrogram *m*: frames in time window *n* : Mel filter banks
Matrix X_i (1 x *m* x *n*)	Matrix X_i (1 x *m* x *n*)
1 x conv3-1;s1 Maxpooling 3x3	LSTM model
1x conv5-1;s1 Maxpooling 3x3	Layers =2
3 Fully Connected layers	Hidden size= 128
OUTPUT: sensor value	OUTPUT: sensor value

Figure 3.3 – Deep neural network configurations of the spectral based methods for sensor value prediction

In the CNN model (Schmidhuber, 2015), the data is fed into a network of two convolution layers with a single channel and kernel size of 3 and 5 respectively for filtering operation to extract local feature maps. Max-pooling is deployed to reduce the dimensionality of feature maps while retaining the vital information. The rectified linear unit activation function is applied to introduce non-linearity into the feature extraction process for each convolution layer, as shown in Figure 3.3. Batch normalization is also applied on each convolution layer. This is followed by 3 fully connected layers with ReLU activation function. Adam optimizer (Kingma et al., 2015) with a learning rate of 0.001 is used as an optimization algorithm.

In the LSTM-RNN model, the data is fed into a network of two LSTM layers with 128 hidden units and a learning rate of 0.001. Adam optimizer is used as an optimization algorithm to update network weights iterative based on training data (Kingma et al., 2015). These hyperpa-

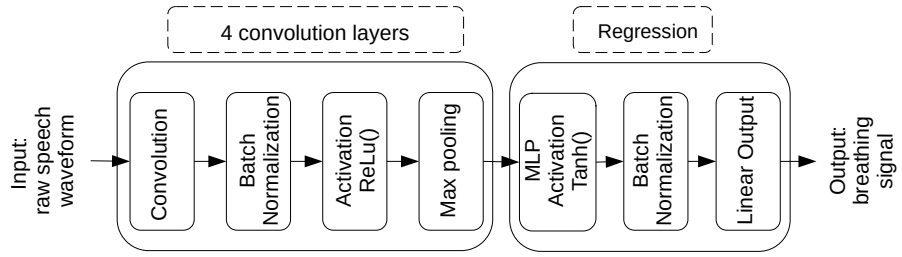


Figure 3.4 – An illustration of the end-to-end CNN model used in raw waveform based methods.

rameters for the network are best chosen for estimation after repeated experimentation.

As estimating breathing pattern from speech using neural networks is a regression problem, we use the following two metrics for evaluation and comparison: Correlation and mean squared error (MSE) of estimated breathing signal and the actual respiratory sensor signal. Also, we compare the breathing parameters derived from the estimated and actual breathing signals. The model that estimates breathing signals with a higher correlation, lesser MSE, and comparable breathing parameters would be considered best for our study.

3.3.2 Raw speech waveform based approach

We used an end-to-end CNN-based model to predict breathing signal values from raw speech waveform as explained in Section 2.2. The CNN consists of four convolution and max-pooling layers followed by a fully connected layer (MLP), and finally, an output layer, as illustrated in Figure 3.4. The number of filters in convolution layers is 128-256-512-512, with kernel sizes of 30-10-4-3 and kernel strides of 10-5-2-1. There are max-pooling layers with strides of 2-3-1-1 and a rectified linear unit (ReLu) as activation function after each layer. The MLP has one hidden layer with 10 units with hyperbolic tangent (Tanh) as activation. The output layer consists of one or two units with linear activation depending on the study. Batch normalization is also applied after each layer. Adam optimizer with learning rate of 0.001 is used for training. The system is implemented using Tensorflow (Abadi et al., 2016).

The input to the system is raw speech waveform, and the output of the neural network is a sample-by-sample prediction of the breathing signal and, when needed, the sensor gradient as well 3.4.2.1. In the latter case, the output of the network is two-dimensional.

3.3.3 Fusion based approach

In this approach, we chose the best model from each of the previously mentioned approaches 3.3.1, 3.3.2, and we fuse the predicted breathing signal values by aligning the two signal and taking the mean of them. We expect the fused estimated breathing signal to be much closer to the actual breathing signal as it has the prospects of both spectral and raw waveform methods.

3.3.4 Regression loss functions

The loss function of a neural network model characterizes how well the model performs over training data. We investigated various loss functions in this study, starting with the standard regression loss functions to customized loss functions based on our problem of estimating breathing signal.

1. **Mean squared error loss function:** It is the most common regression loss function. It is a quadratic loss (L2 loss) and is the sum of squared distances between the target variable and predicted values.

2. **BerHu loss function:** In speech breathing patterns, the breath events are usually a sudden peak (inhalation) followed by a gradual descending curve (exhalation). Thus for the model to estimate breathing patterns, the loss function should be more sensitive to peaks (outliers) and less sensitive for the rest, which can be achieved by using BerHu loss function (Zwald et al., 2012).

$$L_\delta(y, f(x)) = \begin{cases} (|y - f(x)| - 0.5\delta), & \text{if } |y - f(x)| \leq \delta \\ \delta * 0.5 * (y - f(x))^2, & \text{otherwise} \end{cases} \tag{3.3}$$

   We use BerHu loss as loss function for robust regression by integrating the advantages of both the L2 norm and L1 norm, thus penalizing the outliers (peaks) resulting in accelerated optimization of the model for estimating breathing signal.

3. **Correlation loss function:** We use Pearson's correlation coefficient as a measure of similarities between the predicted and actual breathing signals. It is, therefore, reasonable to define a loss function that optimizes this measure directly. We defined a custom Correlation loss function as following:

$$L_{Corr}(y, f(x)) = \frac{1}{1 + r(y, f(x))} - 0.5 \tag{3.4}$$

   where $r(y, f(x))$ is the Pearson's correlation coefficient.

   During training, the Correlation loss is computed by predicting the output signal values for a fixed number of consecutive samples and then calculating the Pearson's correlation coefficient, $r(.)$, between the predicted and the ground truth values. Hence, the time dependency between consecutive samples is taken into account during training.

   We considered the number of consecutive samples, correlation window, as a hyperparameter.

4. **Correlation-MSE loss function:** The custom Correlation loss function removes all the scaling information and focuses on the cycles of the signal. It does not necessarily enforce the actual predicted signal values to get closer to the ground truth. To account for this aspect, we use a combination of Correlation and MSE loss in our training. The custom Correlation-MSE loss function is defined as the following:

$$L_{Corr-MSE}(y, f(x)) = L_{MSE}(y, f(x)) + L_{Corr}(y, f(x)) \tag{3.5}$$

Where $L_{MSE}(y, f(x))$ is the mean squared error loss and $L_{Corr}(y, f(x))$ is the custom Correlation loss defined in Equation 3.4.

We investigate and compare the mean squared error loss function, and Berhu loss function for spectral features based methods, and the mean squared error loss function, customized Correlation, and Correlation-MSE loss functions for raw waveform based methods. The selection of appropriate loss functions has been reported based on repeated experimentation for significant performance improvement.

3.3.5 Hyperparameters for models

The following are the essential parameters to be tuned for the design of our neural network models.

1. **Length of time window:** The window length for each speech input representation, i.e., raw waveform, spectrograms, and log Mel spectrograms, is crucial for estimating the breathing sensor value. We investigate speech inputs of fixed window length of 2, 4, and 8 seconds for spectral representations and 2, 3, and 4 seconds for raw waveform for better estimation.
2. **Length of correlation window:** It is defined as the number of consecutive points in the output of the system that is used to find the correlation between ground truth and predicted values during training. We investigate systems with correlation window sizes of 400, 512, and 1024.
3. **Mapping point of respiratory sensor:** Speech signal of a fixed time window is mapped with respiratory sensor value at the endpoint of the time window to train the models. We investigated mapping with sensor value at the beginning and midpoint of the time window for Philips read speech database and found no significant difference in the estimation performance. Thus, we extended the same endpoint mapping for models for both read speech and conversational speech protocols.

3.4 Experimental Setup

3.4.1 Database and protocols

The study explores the respiratory analysis on the following two protocols: read speech and conversational speech. We use a speech database developed at Philips Research for read speech (Nallanthighal et al., 2019), and for conversational speech, we use the UCL Speech Breath Monitoring (UCL-SBM) database (Schuller et al., 2020).

The Philips read speech database was collected at Philips Research, Eindhoven, The Netherlands in 2019, with the approval of the Internal Committee Biomedical Experiments (ICBE) of Philips Research. The data was collected using the following setup: two respiratory elastic transducer belts over the ribcage under the armpits and around the abdomen at the umbilicus level to measure the changes in the cross-sectional area of ribcage and abdomen at the sample rate of 2 kHz. These belts work on the principle of respiratory inductance plethysmography (RIP). They consist of a sinusoidal wire coil insulated inelastic. The belts' dynamic stretching creates waveforms due to changes in self-inductance and oscillatory frequency of the electronic signal. The electronics convert this change in frequency to a digital respiration waveform where the waveform's amplitude is proportional to the inspired breath volume. Thus the sum of the rib cage and abdomen expansions measured by the respiratory belt transducers is considered as the measure for the breathing signal. Earthworks microphone M23 is used for recording high-quality speech at 48 kHz. The microphone is placed at a distance of one meter from the speaker, and the data collection is conducted in a specialized audio room for noise-free and echo-free recordings. 40 healthy subjects with no respiratory conditions (18 female and 22 male with age group ranging from 21 to 40 years old) are asked to read "The Rainbow Paragraph", a widely used phonetically balanced paragraph (Fairbanks, 1960).

UCL Speech Breath Monitoring (UCL-SBM) database has been introduced in (Schuller et al., 2020). It includes recordings from 49 speakers, which are divided into three non-overlapping subsets; 17 speakers in *Train*, 16 speakers in *Dev*, and 16 speakers in *Test* subset. However, we use a subset of this database: 17 speakers in *Train* subset and 16 speakers in *Dev* subset for training, validating, and testing as the respiratory signals of *Test* subset speakers are not publicly available. For each speaker, a 4 minutes recording of speech with a sampling frequency of 16 kHz is provided. For speakers in *Train* and *Dev* sets, the breathing signal with a sampling frequency of 25 Hz is provided, which amounts to a sequence of 6000 values for each speaker. In this database, the respiratory signal is recorded from one (MLT1132, ADInstruments, Castle Hill, Australia) of the two piezoelectric respiratory belts worn by the subjects. The belt is positioned approximately four centimeters below the collarbone to record chest breathing and produces a linear voltage reading in response to changes in thoracic circumference associated with respiration.

3.4.2 Systems

Our experiments are based on the breathing signal estimation using neural network models in two different protocols: read speech protocol and conversational speech protocol, as described in Section 3.4.1.

In each protocol, performance of neural network models based on spectral features and raw waveform are compared with different cost functions (Section 3.3.4) and evaluated based on metrics of evaluation (Section 3.4.4). A fusion of two estimated breathing signals, each from neural network models based on spectral features and neural network models based on the

raw waveform, is also reported. The same neural network models and system configurations are used for both read speech and conversational speech protocol. System configurations are explained in detail in this section.

3.4.2.1 Spectral features based systems

In spectral features based systems, for the Philips database of read speech protocol, we use 29 subjects for training, 3 subjects for validation, and 8 subjects for testing. Here the speech signal is sampled at 48 kHz, and the breathing signal is downsampled to 100 Hz. For the UCL-SBM database of conversational speech protocol, we use 15 subjects from the subset *Train* for training and the remaining 2 subjects from the *Train* subset for validation. All the 16 subjects of subset *Dev* is used for testing. Here the speech signal is sampled at 16 kHz, and the breathing signal's sampling frequency is 25 Hz. We use MSE and BerHu loss as described in Section 3.3.4 to train our systems.

Respiratory sensor as output: Spectral features of the speech (spectrograms or log Mel spectrogram as described in Section 3.3.1) of a fixed time window (2s, 4s, and 8s) are mapped with the respiratory sensor value at the endpoint of the time window. This is based on our hypothesis that the respiratory sensor value (breathing state) at the end of a time window is dependent on the composition of speech, i.e., linguistic content and prosodic factors in that particular time window. Spectral features of speech and known respiratory sensor values are mapped with a stride of 10ms between windows for Philips database and a stride of 40ms between windows for UCL-SBM database to train neural network models, as defined in Section 3.3.1. These trained models are used to estimate the respiratory sensor values from a target speech signal in real-time to estimate the breathing signal. We compare the performance of the CNN and LSTM-RNN models using the spectral representations and fixed time window.

Table 3.1 – Comparison of window lengths for spectrograms and log Mel spectrograms with MSE loss function for sensor vs speech signal model.

Models	***Window size (s)***	***r***	**MSE**
CNN Model	2	0.26	0.066
log Mel Spectrogram	**4**	**0.472**	**0.034**
	8	0.32	0.030
LSTM RNN Model	2	0.36	0.026
Log Mel Spectrogram	**4**	**0.476**	**0.019**
	8	0.34	0.031
CNN Model	2	0.29	0.096
Spectrogram	4	0.27	0.016
	8	0.24	0.062
LSTM RNN Model	2	0.24	0.082
Spectrogram	4	0.21	0.058
	8	0.23	0.074

The results of Table 3.1 suggests that log Mel spectrogram is a preferred input spectral feature representation of speech signal, and a fixed time window of 4s provides the least mean squared error loss and a high correlation for estimating the breathing signal. With this inference, we investigate all the models based on spectral features approach with input representation as log Mel spectrograms and fixed window length of 4s.

Respiratory sensor and sensor gradient as output: The gradient of the respiratory sensor signal over a fixed time window can be visualized as the net airflow equivalent (inhalation and exhalation) during that fixed time window. We investigate by mapping the spectral features of the speech of a fixed time window with the gradient of the sensor signal over a fixed time window. Log Mel spectrogram and known respiratory sensor gradient values are mapped with a stride of 10ms between windows for Philips database and a stride of 40ms between windows for UCL-SBM database to train deep neural network model so that the model understands the relationship of airflow over a speech utterance in a fixed time window. We used the same CNN and LSTM-RNN model architecture used for sensor and speech mapping models and observed a positive correlation close to 0.2, which explains its relevance but not sufficient to get good performance for estimating the breathing signal. However, by mapping speech signal of a fixed time window to both respiratory sensor value at the endpoint of the time window and respiratory sensor gradient over the fixed time window of 4s, we may achieve a better estimation of breathing pattern and enable the model to learn the breathing state and airflow relationship over a speech utterance. Based on this hypothesis, we investigate by training the models by mapping both the sensor and the sensor gradient for better estimation.

3.4.2.2 Raw speech waveform based systems

In raw speech waveform based systems, for the Philips database of the read speech protocol, similar to the systems using spectral features, we use 29 subjects for training, 3 subjects for validation, and 8 subjects for testing. However, we downsample the speech signal to 16 kHz and the breathing signal to 25 Hz. This is done for reducing the computational time and complexity for training the models based on raw speech waveform. For conversational speech protocol, similar to the systems using spectral features, we use 15 subjects of subset *Train* for training, the remaining 2 subjects of subset *Train* for validation, and 16 subjects of subset *Dev* for testing. Here the speech signal is sampled at 16 kHz, and the breathing signal's sampling frequency is 25 Hz. We train our systems with an input window length of 2s, 3s, and 4s and a correlation window length of 400, 512, and 1024 output samples. We use MSE, Correlation, and Correlation-MSE loss as described in Section 3.3.4 to train our systems.

Respiratory sensor as output: In this set of experiments, similar to the case for spectral features based systems, we map the raw speech waveform of input window length to sensor value at the endpoint of this window. We use a stride of 40 ms between windows to train the models for both Philips and UCL-SBM databases. The result are shown in Tables 3.3 and 3.4. The hyperparameters for the best performance is variable based on the database and loss function. The details for choosing the best systems is explained in Section 3.5.

Respiratory sensor and sensor gradient as output: In this set of experiments, similar to the case for spectral features based systems, we map the raw speech waveform of input window length to both sensor value and the sensor gradient value over the input window length. We use a stride of 40 ms between windows to train the models for both Philips and UCL-SBM databases. We investigate if the gradient information will help our systems to learn the breathing pattern more accurately. The results are shown in Tables 3.3 and 3.4.

3.4.2.3 Fusion of the best systems

As mentioned in Section 3.3.3, we fuse the best system from spectral features based and raw speech waveform based approaches and investigate system performance. We perform this analysis for the systems trained by using only sensor values and systems trained by using both sensor and sensor gradient values on both read speech and conversational speech protocols. In case of Philips database where the output signal sampling frequency is different between the two approaches we downsample the result from the spectral based method to 25 Hz, to be consistent with the output sampling frequency of raw waveform based database. The results are reported in Tables 3.3 and 3.4 and discussed in Section 3.5.

3.4.3 Cross database study

To investigate the generalization ability of the models trained with read speech and conversational speech protocols, we perform a cross database analysis. We consider the following scenarios:

1. Train on Philips read speech database and test on UCL-SBM conversational speech database.
2. Train on UCL-SBM conversational speech database and test on Philips read speech database.

It is interesting to obtain good performance in either of these scenarios which suggest that the models are learning the common aspects of the speech and breathing signal relationship independent of the speech protocol. The results are reported in Tables 3.3 and 3.4 and further discussed in Section 3.5.

3.4.4 Metrics for evaluation

Estimating breathing signal from speech is a regression problem. The following metrics are used to evaluate the estimation of the predicted breathing signal: Pearson's correlation coefficient, mean squared error, and breathing parameters.

A higher Pearson's correlation ensures that the prediction follows the trend of the actual signal and a low mean squared error ensures that the prediction is in a dynamic range of the actual signal. Breathing parameters as described in Section 3.2, derived from actual and estimated breathing signal should be comparable with minimum error and high breath event sensitivity. Thus a model which can estimate the breathing signal with higher Pearson's correlation coefficient, lower mean squared error with respect to the actual signal and comparable breathing parameters can be considered best.

3.5 Results

The results for the experiments explained in Section 3.4 are presented here. The systems are trained using the Philips database of the read speech protocol and UCL-SBM database of the conversational speech protocol. We further performed a cross database analysis to investigate the generalization abilities of our systems.

We trained our systems with the mentioned loss functions and hyperparameters as discussed in Sections 3.3.4 and 3.3.5. In spectral based approaches, we found out that MSE and BerHu loss functions are performing better than correlation based loss functions. As mentioned in Table 3.1, we observed that using log Mel spectrogram with input window length of 4s performs better compared to other combinations. In the rest of this chapter, we report on the

spectral based methods trained with log Mel spectrogram with a fixed window length of 4s and MSE and BerHu loss. We similarly trained our systems with different loss functions for the raw waveform based approach and observed that the MSE, Correlation, and Correlation-MSE loss perform better. Therefore, we report the systems using these three loss functions. We trained our system with an input window of 2s, 3s, and 4s, and the correlation window size of 400, 512, and 1024 samples. The best performing system is different for Philips and UCL-SBM database and for different loss functions. In each case, we chose the system with the best performance in terms of correlation among those trained with only respiratory sensor values as an output and reported their performance. Table 3.2 summarizes the hyperparameters of the reported systems for the raw waveform based methods on both Philips and UCL-SBM databases.

Table 3.2 – The hyperparameters of the reported systems for the raw waveform based approach. They have been chosen based on the Pearson's correlation coefficient of the systems trained using only respiratory sensor values as output.

Database	Loss function	Input window length (s)	Correlation window length (samples)
Philips	MSE	3	–
	Corr	4	400
	Corr-MSE	4	400
UCL-SBM	MSE	3	–
	Corr	2	1024
	Corr-MSE	3	1024

We are also reporting breathing parameters, including breathing rate error, breathing event sensitivity, and tidal volume error rate in each case. The important aspect while computing breathing parameters from the actual and estimated breathing signal is fixing the look-ahead window for computing the peaks in the signals using automatic multiscale peak detection (AMPD) algorithm. We fix this look-ahead window to 2s which constitutes to 200 samples when the sampling rate of the breathing signal is 100 Hz and to 50 samples when the sampling rate is 25 Hz. This 2s window is selected based on our observation that the minimum gap between two inhalation breath events is more than 2s. We use the same look-ahead window of 2s for both true breathing signal and estimated breathing signal for a fair comparison. We observe that the estimated breathing rate is usually higher than the true breathing rate in Tables 3.3, 3.4, 3.5, and 3.6. This is due to the occurrence of intermittent peaks in the estimated breathing signal obtained from neural network models. Also, the estimated breathing signals usually are noisier than the true signal and there is high probability for the peak detection algorithm to find a false peak.

Table 3.3 – Philips database (read speech). The abbreviations used for breathing parameters: BR=Breathing Rate, BE= Breathing Events, and TV=Tidal Volume.

Models	Loss Function	*r*	MSE	Breathing Parameters				
				Breathing Rate			Breath Event	Tidal Volume
				prediction (breaths/min)	true (breaths/min)	error (%)	Sensitivity	error (%)
Spectral Based Approach								
I/P: log Mel spec O/P: Sensor Architecture: RNN	MSE BerHu	0.476 0.482	0.019 0.039	10.42 10.98	9.84 9.84	5.89% 11.58%	0.916 0.902	12.11% 16.24%
I/P: log Mel spec O/P: sensor & sensor gradient Architecture: RNN	MSE BerHu	0.452 0.463	0.021 0.019	11.03 11.80	9.84 9.84	12.09% 19.91%	0.882 0.842	11.62% 14.72%
I/P: log Mel spec O/P: sensor Architecture: CNN	MSE BerHu	0.472 0.462	0.034 0.042	10.85 11.78	9.84 9.84	10.26% 19.71%	0.896 0.821	16.22% 18.84%
I/P: log Mel spec O/P: sensor & sensor gradient Architecture: CNN	MSE BerHu	0.437 0.413	0.051 0.063	11.52 12.10	9.84 9.84	17.68% 22.96%	0.847 0.811	19.86% 20.46%
Raw Waveform Based Approach								
I/P: raw waveform O/P: sensor Architecture: CNN	MSE Corr Corr-MSE	0.47 0.534 0.502	0.0699 1.8627 0.0616	12.00 11.81 11.24	9.90 9.90 9.90	21.29% 19.35% 13.55%	0.929 0.942 0.897	18.47% 14.4% 26.65%
I/P: raw waveform O/P: sensor & sensor gradient Architecture: CNN	MSE Corr Corr-MSE	0.45 0.449 0.447	0.068 0.8857 0.0874	12.51 12.45 11.30	9.90 9.90 9.90	26.45% 25.81% 14.19%	0.903 0.955 0.890	30.5% 45.69% 8.25%
Fusion Based Approach								
Model_1: spectral based Model_2: raw waveform based O/P: sensor	MSE Corr	0.562	1.024	10.68	9.84	8.53%	0.908	10.21%
Model_1: spectral based Model_2: raw waveform based O/P: sensor & sensor gradient	MSE Corr-MSE	0.480	0.071	11.18	9.84	13.61%	0.882	14.42%

3.5.1 Philips database study

Table 3.3 presents the performance of the systems trained and tested on the Philips database for the three approaches.

Looking into the systems trained with only sensor values, it is evident that the raw waveform based methods, on average, yields a slightly higher Pearson's correlation coefficient and a higher MSE. The MSE for systems trained with Correlation loss, however, is much higher than other methods. The spectral based methods yield lower breathing rate and tidal volume error than the raw waveform based methods; however, the breath events sensitivity is also slightly lower. In general, both spectral features based and raw waveform based approaches perform similarly, considering all the mentioned evaluation metrics. The spectral features based method with the highest Pearson's correlation coefficient is an RNN model trained with BerHu loss and $r = 0.482$. Even though comparable to other systems, the MSE and breathing parameters for this model are not the best. The next highest r value belongs to the RNN system trained with MSE loss function, which has lower MSE and performs better in terms of breathing parameters. Considering all the evaluation metrics, we chose the RNN model trained with MSE loss as the best model to be used for the fusion approach. In the case of

raw waveform based methods, the system with the highest Pearson's correlation coefficient is trained with Correlation loss function with $r = 0.534$. This is the highest correlation among all the trained systems, not considering the fusion systems. The MSE for the same model is also the highest, with a value of 1.8627. It can be due to the nature of the defined Correlation loss function, which removes any scaling from the data and only focuses on the signal's repetitive temporal aspect. Unlike MSE loss functions, correlation based loss functions do not decrease the distance between actual and predicted samples. The breathing parameters for this model are comparable or sometimes better than other systems. The reason for this can be because the peak detection algorithm is not sensitive to the actual peak value of the signal. Similarly, the tidal volume computation focuses on relative behavior of minimum and maximum value of the signal around a peak, but not their actual value. By combining Correlation and MSE loss, we acquire a system that performs comparable to the system with only correlation loss, but the MSE is much lower, and the breathing parameters, if not the best, are comparable to other systems. Adding MSE to the loss functions obligates the actual predicted values to get closer to the true values.

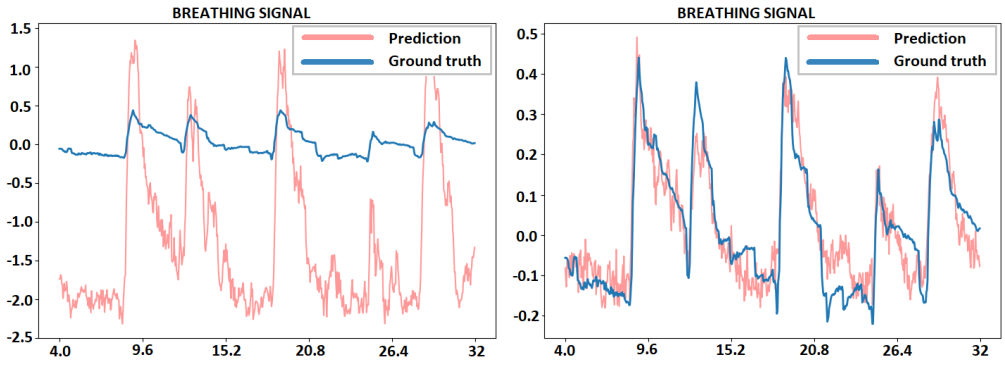
Looking into the systems trained with both sensor and sensor gradient, we observe a slight decrease in the Pearson's correlation coefficient of all the systems and increase in the MSE in most of them. They behave differently with regard to breathing parameters however, it can be noted that overall we obtain a less performance using both sensor and sensor gradient values. For the systems trained with the spectral based approach, considering all the evaluation metrics, the system with the best performance is the LSTM-RNN trained with MSE loss. For raw waveform based approach, considering all the evaluation metrics, the system with the best performance is the CNN trained with Correlation loss for sensor output models and Correlation-MSE loss for sensor and sensor gradient output models. These two systems from each approach are used for the fusion based approach. The fusion system's performance obtained by aligning and averaging the predicted values for the chosen best systems from each approach is presented in the last section of Table 3.3. It can be seen that by only fusing the two systems in the output level, we can gain a boost in the performance of our system specially in terms of correlation. We benefit from information learned by the two individual systems and obtain a system that performs better. This can be beneficial as we can train smaller and less complicated systems that are easier to train and still benefit from them.

Another point that can be seen in Table 3.3 is regarding the breathing rate for the ground truth and predicted values. The true breathing rate for the labels reported in the spectral based approach (9.84 breaths/minute) is slightly different from the raw waveform based approach (9.90 breaths/minute). The reason is that in our implementation of the spectral based methods, the first sample of breathing signal mapped to spectral features is the endpoint of the first 4s window. Thus, the first 4s at the beginning of the breathing signal is dropped for each subject and this slight difference in the length of the breathing signal results in slight difference of breathing rates in the two approaches.

3.5.2 UCL-SBM database study

We performed a similar investigation using the UCL-SBM database. Table 3.4 presents the results for the spectral features, raw waveform, and fusion based methods. When looking to the systems trained with only sensor values, compared to the systems trained on the Philips database, on average, we obtain slightly lower Pearson's correlation coefficient, higher MSE, and higher error in the breathing parameters. The Pearson's correlation coefficient for the raw waveform based methods are very similar to those of the spectral based methods. The MSE for the system trained with the Correlation loss is much higher than other systems, while the obtained Pearson's correlation coefficient is also the highest. Figure 3.5 shows an example of the predicted and ground truth values for the breathing signal when trained using Correlation and Correlation-MSE loss. From Figure 3.5 the reason for such a high MSE when using Correlation loss is evident. The system is not able to predict the appropriate dynamic range of the output and therefore, even though the predicted signal is following the same trend as the ground truth, their actual values are far from each other. For the systems trained with the spectral based approach, considering the evaluation metrics, the system with the best performance is the LSTM-RNN trained with MSE loss. For raw waveform based approach the system with the best performance is the CNN trained with Correlation loss for sensor output models and Correlation-MSE loss for sensor and sensor gradient output models. These two systems from each approach are used for the fusion.

Looking into the systems trained with both sensor and sensor gradient values we again do not see much improvement. From the last section of Table 3.4, fusing the best systems from two approaches, we obtain a system that performs better than the individual systems similar to Philips database.



(a) **(b)**

Figure 3.5 – The predicted and ground truth for the breathing signal for a raw waveform based method trained on the UCL-SBM database using (a) Correlation and (b) Correlation-MSE loss functions.

Table 3.4 – UCL-SBM database (conversational speech). The abbreviations used for breathing parameters: BR=Breathing Rate, BE= Breathing Events, and TV=Tidal Volume.

Models	Loss Function	*r*	MSE	Breathing Parameters				
				Breathing Rate			Breath Event	Tidal Volume
				prediction (breaths/min)	true (breaths/min)	error (%)	Sensitivity	error (%)
Spectral Based Approach								
I/P: log Mel spec	MSE	0.482	0.039	10.44	9.35	11.65%	0.908	13.42%
O/P: sensor	BerHu	0.448	0.018	10.71	9.35	14.42%	0.882	11.68%
Architecture: RNN								
I/P: log Mel spec	MSE	0.463	0.019	10.42	9.35	11.44%	0.871	08.74%
O/P: sensor & sensor gradient	BerHu	0.427	0.016	10.84	9.35	15.93%	0.840	10.55%
Architecture: RNN								
I/P: log Mel spec	MSE	0.437	0.042	11.11	9.35	18.82%	0.841	19.29%
O/P: sensor	BerHu	0.460	0.042	10.57	9.35	13.04%	0.864	18.62%
Architecture: CNN								
I/P: log Mel spec	MSE	0.411	0.036	10.62	9.35	12.24%	0.810	21.72%
O/P: sensor & sensor gradient	BerHu	0.413	0.063	11.44	9.35	22.31%	0.822	24.82%
Architecture: CNN								
Raw Waveform Based Approach								
I/P: raw waveform	MSE	0.411	0.0263	10.45	9.39	11.31%	0.887	28.13%
O/P: sensor	Corr	0.490	2.107	13.12	9.39	39.77%	0.982	12.94%
Architecture: CNN	Corr-MSE	0.463	0.0253	11.55	9.39	22.96%	0.933	18.25%
I/P: raw waveform	MSE	0.406	0.0268	09.20	9.39	02.00%	0.797	23.15%
O/P: sensor & sensor gradient	Corr	0.470	2.464	12.98	9.39	38.27%	0.957	08.36%
Architecture: CNN	Corr-MSE	0.459	0.0253	11.61	9.39	23.63%	0.928	17.92%
Fusion Based Approach								
Model_1: spectral based RNN	MSE							
Model_2: raw waveform based	Corr	0.512	1.822	11.24	9.35	20.21%	0.942	12.33%
O/P: sensor								
Model_1: spectral based RNN	MSE							
Model_2: raw waveform based	Corr-MSE	0.466	0.022	10.64	9.35	13.7%	0.862	14.44%
O/P: sensor & sensor gradient								

3.5.3 Cross database study

We performed a cross database investigation for the systems trained with only sensor values. We used the systems trained on the Philips database to predict the breathing signal on UCL-SBM database and calculated the performance. Table 3.5 shows the result of our investigation. It can be seen that the spectral based methods are performing better compared to the raw waveform based methods, however their performance decreases compared to the performance when tested on the Philips database. We can gain even a better performance by fusing the systems from the two approaches and obtain a system with a correlation of 0.398, MSE of 0.042, and comparable breathing parameters.

Table 3.6 shows the result of testing the systems on the Philips database while trained on the UCL-SBM database. Once again, we observe a better performance for the spectral based methods than the raw waveform based methods. The best system performs with a correlation of 0.364, MSE of 0.049, and comparable breathing parameters in the spectral based approach. The performance of the raw waveform based methods decreases drastically in this case. We observe that fusing the two systems seems to perform slightly worse than only the spectral based model.

Table 3.5 – Train on Philips (read speech) database and test on UCL-SBM (conversational speech) database. The abbreviations used for breathing parameters: BR=Breathing Rate, BE= Breathing Events, and TV=Tidal Volume.

Models	Loss Function	*r*	MSE	Breathing Parameters				
				Breathing Rate			Breath Event	Tidal Volume
				prediction (breaths/min)	true (breaths/min)	error (%)	Sensitivity	error (%)
Spectral Based Approach	MSE	0.372	0.039	10.39	9.35	11.12%	0.872	15.64%
	BerHu	0.344	0.031	11.04	9.35	18.07%	0.820	14.20%
Raw Waveform Based Approach	MSE	0.353	0.0457	11.83	9.39	25.96%	0.895	29.48%
	Corr	0.284	2.2949	12.70	9.39	35.27%	0.933	18.57%
	Corr-MSE	0.299	0.0562	11.73	9.39	24.96%	0.867	3.20%
Fusion Based Approach	MSE,MSE	0.398	0.042	10.44	9.35	11.65%	0.868	14.62%

Table 3.6 – Train on UCL-SBM (conversational speech) database and test on Philips (read speech) database. The abbreviations used for breathing parameters: BR=Breathing Rate, BE= Breathing Events, and TV=Tidal Volume.

Models	Loss Function	*r*	MSE	Breathing Parameters				
				Breathing Rate			Breath Event	Tidal Volume
				prediction (breaths/min)	true (breaths/min)	error (%)	Sensitivity	error (%)
Spectral Based Approach	MSE	0.364	0.049	10.42	9.84	5.89%	0.916	14.42%
	BerHu	0.331	0.071	11.24	9.84	14.22%	0.853	21.62%
Raw Waveform Based Approach	MSE	0.129	0.0727	9.13	9.9	7.74%	0.684	9.52%
	Corr	0.217	2.5687	13.15	9.9	32.9%	0.942	8.76%
	Corr-MSE	0.070	0.0922	12.32	9.9	24.52%	0.923	1.27%
Fusion Based Approach	MSE,Corr	0.347	2.346	10.66	9.84	8.3%	0.898	11.22%

3.6 Analysis of Proposed Approaches

This section presents an analysis of the studied approaches.

3.6.1 MAE loss function

In this chapter, we have investigated different loss functions (see Section 3.3.4). In the speech enhancement literature, the mean absolute error (MAE) loss function has also been proposed for regression (Qi et al., 2020). We conducted an analysis study using the MAE loss function on the Philips database. Table 3.7 presents the results of the study. It can be observed that MAE loss tends to yield comparable to those trained with MSE loss, see Table 3.3, both in terms of correlation coefficients and breathing parameter estimation. We do not observe a distinctive advantage as in the case of speech enhancement studies.

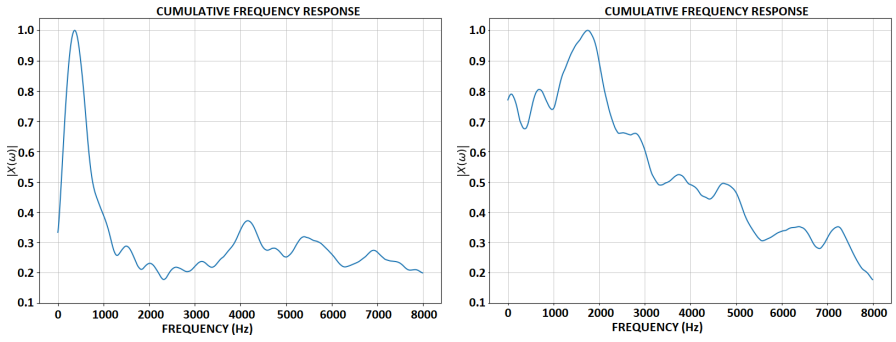
3.6.2 Analysis of raw waveform based approach

In the experimental studies, we observed a trend that spectral based approach typically yields better performance than the raw waveform based approach, particularly in terms of MSE

Table 3.7 – MAE loss function for Philips database (read speech). The abbreviations used for breathing parameters: BR=Breathing Rate, BE= Breathing Events, and TV=Tidal Volume.

Models	Loss Function	*r*	MSE	Breathing Parameters				
				Breathing Rate			Breath Event	Tidal Volume
				prediction (breaths/min)	true (breaths/min)	error (%)	Sensitivity	error (%)
Spectral Based Approach								
RNN (I/P: log Mel spec, O/P: sensor)	MAE	0.448	0.027	11.15	9.84	13.31%	0.864	14.02%
RNN (I/P: log Mel spec, O/P: sensor & sensor gradient)	MAE	0.411	0.029	12.11	9.84	23.06%	0.862	17.27%
CNN (I/P: log Mel spec, O/P: sensor)	MAE	0.451	0.047	11.55	9.84	17.37%	0.844	23.24%
CNN (I/P: log Mel spec, O/P: sensor & sensor gradient)	MAE	0.419	0.081	12.03	9.84	22.25%	0.872	24.12%
Raw Waveform Based Approach								
CNN (I/P: raw waveform, O/P: sensor)	MAE	0.489	0.0572	11.68	9.90	18.06%	0.897	12.72%
CNN (I/P: raw waveform, O/P: sensor)	Corr-MAE	0.507	0.0592	11.94	9.90	20.65%	0.929	20.09%
CNN (I/P: raw waveform, O/P: sensor & sensor gradient)	MAE	0.431	0.063	11.3	9.90	14.19%	0.877	47.39%
CNN (I/P: raw waveform, O/P: sensor & sensor gradient)	Corr-MAE	0.486	0.0548	10.66	9.90	7.74%	0.858	12.01%

and correlation. This is more evident in the cross database investigation. To understand the reason behind that, we analysed the first convolution layer filers by computing the cumulative frequency response. Figure 3.6 shows the cumulative frequency response of the first convolution layer for CNN trained on Philips database and UCL-SBM database. It can be seen that in both cases, the filters are giving emphasis to low but different frequency regions. This somewhat explains the drop in *r* and MSE in the cross database study. The cumulative frequency responses also indicate the difference in performances between the spectral based approach and the raw waveform based approach. More precisely, log Mel frequency filter bank energies tend to characterize the spectral envelope covering the entire bandwidth. Raw waveform based approach is giving emphasis to the selective frequency region. To ascertain whether this difference impacts the performance, we trained the spectral based approach systems with the first 20 log Mel filter bank energy as input with MSE loss. The mid-frequency of the 20th filter is 1949.99 Hz. This is close to the frequency region the CNNs in the raw waveform based approaches are emphasizing.



(a) **(b)**

Figure 3.6 – The cumulative frequency response of the kernels for the first layer of the CNN model trained on (a) Philips and (b) UCL-SBM databases, using Correlation-MSE loss function.

Table 3.8 compares the performance achieved with the first 20 log Mel filter bank energy as input (denoted as RNN-20 and CNN-20) with 40 filterbank energies as input (denoted as

RNN-40 and CNN-40, results taken from Table 3.3 Line 1 and Line 5). We can observe that there is a clear drop in performance in both systems in terms of *r* and MSE. In the case of RNN architecture, there is a clear drop in performance in terms of breathing rate error, breathing event sensitivity, and tidal volume error. In the case of CNN architecture, there is not much change in breathing parameter estimation. This indicates that the spectral region modeled has an impact on the performance.

Table 3.8 – Comparison between first 20 log Mel filterbank energies and 40 log Mel filterbank energies as input. The loss function used for all the systems is MSE. The abbreviations used for breathing parameters: BR=Breathing Rate, BE= Breathing Events, and TV=Tidal Volume.

Models	***r***	**MSE**	**Breathing Parameters**		
			BR error (%)	**BE Sensitivity**	**TV error (%)**
RNN-20	0.415	0.078	12.8%	0.882	17.2%
RNN-40	0.476	0.019	5.89%	0.916	12.11%
CNN-20	0.423	0.096	12.11%	0.858	15.6%
CNN-40	0.472	0.034	10.26%	0.896	16.22%

It is possible to guide the raw waveform based approach to model spectral envelop related information, similar to the spectral features based approach. In phone classification studies, it has been consistently demonstrated that the raw waveform based approach is able to capture short-term spectral envelop information (Palaz et al., 2019; Muckenhirn et al., 2019). So, we could first pre-train the CNN on phone classification task and then adapt it for breathing signal estimation.

3.6.3 Comparison to other approaches

Sensing breathing signal estimation from speech signal is a relatively new problem. In that direction, the spectral based approaches dealt in this study were the first to be explored (Nallanthighal et al., 2019; Nallanthighal et al., 2020). The Interspeech 2020 ComParE challenge devoted a sub-challenge on breathing signal estimation from speech signal (Schuller et al., 2020). The new methods development has happened in parallel with our work. A fair comparison to those methods based on the different metrics used in this study is not feasible. The reason being that the sub-challenge compared the systems only based on *r*. Nevertheless, for the sake of completeness, we provide a comparison between the approaches investigated in this paper and the ComParE 2020 sub-challenge baseline and other systems developed as part of the challenge (Markitantov et al., 2020; Mendonça et al., 2020).

Table 3.9 provides the performance of various other deep learning techniques explored in this challenge, including the baseline paper (Schuller et al., 2020), winner of the challenge (Markitantov et al., 2020) as well as our proposed systems, which were tested on the *Test* subset of the UCL-SBM database. It is worth mentioning that our proposed systems in the table are

Table 3.9 – Pearson's correlation coefficient (r) reported on the *Dev* set and the *Test* set. For the sake of clarity, our systems are denoted in the following format: *ANN type_input type_loss function*

	Dev	***Test***
	r	***r***
ComParE 2020 Breathing sub-challenge Baselines (Schuller et al., 2020)		
OPENSMILE: COMPARE functionals+SVM	0.244	0.442
OPENXBOW: COMPARE BoAW+SVM	0.226	0.366
End2End: CNN+LSTM RNN	0.507	0.731
Proposed Systems by Markitantov et al. in (Markitantov et al., 2020)		
1D CNN + LSTM (Raw signal)	0.607	0.744
ResNet18 + GRU (128 log Mel)	0.580	0.734
Fusion	**0.640**	**0.763**
Proposed Systems by Mendonça et al. in (Mendonça et al., 2020)		
BiLSTM Original	0.507	0.720
Our Proposed Systems		
CNN_Raw_MSE (2s)(scaling)	0.519	—
CNN_Raw_Corr (4s, 1024)(scaling)	0.514	—
CNN_Raw_Corr-MSE (4s, 512)(scaling)	0.532	0.628
CNN_Raw_Corr-MSE (4s, 512)	0.476	0.636
CNN_Spec_MSE	0.472	0.452
LSTM-RNN_Spec_MSE	0.448	—
Fusion_Raw	**0.552**	0.656
Fusion_Raw_Spec	0.541	**0.707**

coming from different runs during the ComParE challenge. So they are not exactly the same systems presented in Section 3.5.2. In the following, we denote our systems in this format: *ANN type_input type_loss function*. (2s) refers to 2 seconds long input. (4s, 1024) refers to 4 seconds long input with 1024 correlation window size. The system noted with (scaling) refers to the system where a mean subtraction and scaling between -1 and 1 is applied on the output of the CNN. Fusion_Raw refers to the system where the output of CNN_Raw_MSE and CNN_Raw_Corr are aligned through cross correlation and are averaged. Fusion_Raw_Spec refers to the system wherein the LSTM-RNN_Spec_MSE and CNN_Raw_Corr-MSE are combined with the same method as mentioned before. On the *Dev* set, our systems outperform low level descriptor based systems and bag-of-audio-words based systems. Raw waveform based, CNN_Spec_MSE, Fusion-Raw, and Fusion_Raw_Spec yield performance competitive to the baseline CNN+LSTM RNN system. On the *Test* set, our approach Fusion_Raw_Spec is comparable to the end-to-end baseline and the BiLSTM system proposed in (Mendonça et al., 2020) and inferior to the winner of the challenge (Markitantov et al., 2020). It is worth mentioning that, in the *Test* set protocol, the neural networks were trained with training and development data. As a result, r is higher than the development set. As we have already

observed in the different studies presented earlier, r by itself is not a complete indicator of breathing parameter estimation.

3.6.4 Breathing signal prediction evaluation measures

As an end goal, the speech signal based breathing parameter estimation can be regarded as a non-intrusive "instrument" for measuring breathing parameters. In other words, from the predicted signal we should be able to measure well breathing parameters such as breathing rate and tidal volume for application purposes. The neural network training and evaluation does not consider these aspects. So, a question that arises is: whether the evaluation measures used to evaluate the breathing signal prediction models are sufficient indicators for reliable breathing parameter estimation?

To address this question we chose the best performing systems from the spectral based and raw waveform based approaches when only sensor output is predicted and the models trained on the UCL-SBM database. We study these models in (a) single database study, where the models are trained and evaluated on the UCL-SBM database, and (b) cross database study, where the models are trained on the UCL-SBM database and tested on the whole Philips database instead of only *Test* subset of the Philips database. In the following, we denote our systems in this format: *ANN type_input type_loss* function.

Table 3.10 presents the result of the single database study where the models are trained and tested on the UCL-SBM database. They are taken from Table 3.4 and repeated here for simplicity.

Table 3.10 – Train and test on UCL-SBM (conversational speech) database. The results are taken from Table 3.4. The abbreviations used for breathing parameters: BR=Breathing Rate, BE=Breathing Events, and TV=Tidal Volume.

loss Functions	**r**	**MSE**	**Breathing Parameters**		
			BR error (%)	**BE Sensitivity**	**TV error (%)**
Spectral Based Approach					
MSE	0.482	0.039	11.65%	0.908	13.42%
BerHu	0.448	0.018	14.42%	0.882	11.68%
Raw Waveform Based Approach					
MSE	0.411	0.0263	11.31%	0.887	28.13%
Corr	0.49	2.1068	39.77%	0.982	12.94%
Corr-MSE	0.463	0.0253	22.96%	0.933	18.25%

It can be seen that the systems trained with both approaches are performing closely in terms of correlation. This performance is comparable to the ComParE challenge best baseline system performance reported on the *Dev* set (Schuller et al., 2020). It can be observed that other

Table 3.11 – Train on UCL-SBM (conversational speech) database and test on all Philips (read speech) database. The abbreviations used for breathing parameters: BR=Breathing Rate, BE=Breathing Events, and TV=Tidal Volume.

loss Functions	r	MSE	Breathing Parameters		
			BR error (%)	BE Sensitivity	TV error (%)
Spectral Based Approach					
MSE	0.324	0.072	13.22%	0.862	11.56%
BerHu	0.292	0.108	16.82%	0.804	17.84%
Raw Waveform Based Approach					
MSE	0.187	0.0737	10.91%	0.715	4.79%
Corr	0.298	1.9923	33.45%	0.993	5.14%
Corr-MSE	0.137	0.0928	19.81%	0.886	15.3%

metrics calculated for the systems are largely different. For example, the CNN_raw_Corr system with the correlation of 0.49 has a very high MSE and breathing rate error compared to all the other systems, but the tidal volume error for it is among the lowest. If we consider the systems trained with MSE loss, i.e. CNN_Raw_MSE, and LSTM-RNN_Spec_MSE, we observe that they both yield similar error for breathing rate (11.31% vs 11.65%) and MSE (0.026 vs 0.039), while the correlation and tidal volume error for them are very different. Furthermore, if we compare CNN_Raw_Corr-MSE and LSTM-RNN_Spec_BerHu with similar correlation and MSE, we observe a large difference in breathing parameters, around 8% for breathing rate error and 6% for tidal volume error.

Table 3.11 shows the performance of the systems trained on UCL-SBM database when tested on the whole 40 subjects of the Philips database.

It can be seen that the systems corresponding to the spectral based approach are yielding a higher correlation compared to the systems corresponding to the raw waveform based approach, but they yield higher tidal volume error. Furthermore, although the correlation values are considerably low compared to the single database study, the breathing rate error, breath event sensitivity and tidal volume error are in similar ranges. Interestingly, the raw waveform based approach with MSE yields one of the lowest tidal volume error, despite having a very low correlation.

It is evident from these results that Pearson's correlation coefficient and MSE measures are not good indicators of breathing rate error, breath event sensitivity and tidal volume error.

3.7 Conclusion

Speech and respiration are closely related and therefore, it may be possible to use speech analysis for the sensing of the respiratory health of the speaker. In this chapter, we studied this interdependence and addressed the following challenges to establish respiratory analysis of speech as a practical or clinical application.

1. We explored the possibility of estimating breathing signal from the speech in two approaches: spectral analysis and raw waveform analysis, and established that the breathing signal could be reliably estimated directly from speech. We found that fusing the estimated breathing signals obtained from the best models trained on these two approaches is closer to the actual breathing signal.

2. To make the study more robust and reliable for practical or clinical purposes, we studied two significant protocols: read speech and conversational (or spontaneous) speech. We performed individual database studies and cross database studies. The cross database studies yielded promising performances in terms of breathing parameter estimation. This suggests that these methods can be applied independent of databases. However, note that on individual databases results, the estimation is better than cross database.

3. For evaluating the estimated breathing signal from the neural network models, apart from the standard evaluation metrics for the regression problem: mean squared error and correlation with reference to actual breathing signal measured through respiratory inductive belts, we compared the breathing parameters like breathing rate, tidal volume equivalent, and breath event sensitivity. An extensive comparison in terms of these metrics was done in each protocol and reported. These parameters help in establishing the credibility for the practical application of the proposed methods.

4. The extensive breathing signal prediction and breathing parameter estimation studies conducted in this work show that high Pearson's correlation measure or low MSE not necessarily indicates better breathing parameter estimation. Further investigation is needed to develop neural network training loss functions and evaluation measures that take into consideration breathing parameter estimation.

Estimating breathing signal and parameters from the speech signal is an unobtrusive and potentially cost-effective option for long-term breathing monitoring in telehealth care applications. This technology could facilitate continuous breathing activity monitoring aiding a more thorough and adequate assessment for early recognition of abnormal breathing syndromes.

4 Applications of Breathing Pattern Estimation Networks

4.1 Introduction

The respiratory system plays a crucial role in speech production, and changes in breathing patterns can be indicative of various physiological and emotional states. Estimating breathing pattern from speech signals opens up new avenues for non-invasive monitoring of health conditions and can be potentially used in improving available speech technologies. In this chapter, we explore three diverse use cases of breathing pattern estimation (BPE) networks that we developed in Chapter 3. We chose from the raw waveform based CNN models that were trained in Chapter 3 on both Philips and UCL-SBM databases. These models were used without fine-tuning as feature extractors (see Section 2.1.1.2) in a machine learning framework for the following use cases:

- Detecting COVID-19 infection from speech signals: In this study, we distinguish between COVID-19 positive and negative cases. It is formulated as a binary classification problem. The details of the study are presented in Section 4.2.
- Analysing the breathing pattern information in synthetic speech: In this study, we investigate if synthetic speech carries similar breathing pattern information as natural speech. Towards this goal, we use a database of synthetic speech, designed for developing presentation attack detection methods for speaker verification systems. We develop models to distinguish between synthetic speech and genuine speech as a binary classification problem. The details are presented in Section 4.3.
- Continuous emotion recognition from speech signals: The arousal and valence in speech recordings is estimated in this study in a regression framework. The details are presented in Section 4.4.

We conclude the chapter in Section 4.5. The material presented in this chapter was originally published in modified form in (Mostaani, Prasad, et al., 2022; Mostaani and Magimai.-Doss, 2022; Yadav et al., 2022).

4.2 COVID-19 Detection

Corona virus disease 2019 (COVID-19), caused by severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) is primarily a respiratory infection, which has affected the lives of millions of people all over the world. The World Health Organization (WHO) has announced COVID-19 a pandemic on March 2020 (*WHO Coronavirus* 2021).

To detect COVID-19, several diagnostic routines based on collecting saliva or blood from the patients have been effective. However, these tests are slow and take considerable time to produce results. Cough sounds and speech based diagnosis of COVID-19 has gained interest owing to the ease of recording the signals. A database of cough sounds obtained over more than 20,000 participants, from a range of age groups, gender, ethnicity and COVID-19 status, has been collected to facilitate detection of the virus using audio signals (Orlandic et al., 2021). Similar efforts were taking place in the research community (Sharma et al., 2020; Brown et al., 2020; Han et al., 2021). In the speech community, two challenges as part of Interspeech 2021, namely, Interspeech 2021 ComParE challenge (Schuller et al., 2021) and DiCOVA challenge (Muguli et al., 2021) have been organized in that direction.

In one of the earliest studies, spectral parameters such as, spectral centroid, spectral roll-off, and zero crossing rate along with Mel frequency cepstral coefficients (MFCCs) and functionals were modeled in a recurrent neural network (RNN) based and long short-term memory (LSTM) based framework to detect the presence of COVID-19 (Hassan et al., 2020). The system was found to yield a higher classification accuracy for cough and breathing sounds compared to speech. In Interspeech 2021 ComParE challenge, openSMILE features were found to yield better detection when compared to deep neural network based systems on the COVID-19 Speech sub-challenge, while on the COVID-19 Cough sub-challenge, data augmentation together with transfer learning, using pre-trained audio networks was found to yield better detection. (Klumpp et al., 2021) studied the phonetic patterns in COVID-19 speech using deep acoustic model. They observed that the distinct patterns found can not be solely attributed to COVID-19. In (Harvill et al., 2021), it was found that modeling of features obtained from autoregressive predictive coding neural network together with data augmentation improves cough based COVID-19 detection. Other directions include, investigation of auditory motivated features (Das et al., 2021), combination of different spectral feature representations (Ritwik et al., 2021) and modeling of breathing pattern information in cough (Deshpande et al., 2021b).

In recent years, neural network based methods have emerged which can learn information in a task dependent manner from raw speech waveform directly (Palaz et al., 2013; Trigeorgis et al., 2016; Muckenhirn et al., 2018; Muckenhirn et al., 2019). We developed a neural network based method to estimate breathing pattern from speech signals in Chapter 3. In this study, we question: whether embeddings of such pre-trained neural networks without any form of adaptation can be effectively employed for COVID-19 detection? If successful, such methods can potentially serve as alternate means of finding representations that discriminate between

COVID and non-COVID speech, while providing some form of explainability through the tasks on which those networks are trained. Towards this goal, we investigate modeling of embeddings learned by neural networks trained (a) to classify phones and (b) to estimate breathing patterns, and compare them against modeling of knowledge-based paralinguistic features, namely, ComParE low level descriptors (LLDs) (see Section 2.1.1.1) which have been found useful for COVID-19 detection (Schuller et al., 2021; Avila et al., 2021; Södergren et al., 2021). We also analyse the top ranking LLDs and relate them to the information captured by the raw waveform neural networks. This study has been done as part of the second DiCOVA (DiCOVA-II) challenge (*Second DiCOVA challenge* 2021).

4.2.1 Proposed method

Figure 4.1 illustrates the proposed neural embeddings based approach. This pipeline follows the modular approach presented in Section 2.1 in which, frame-level neural embeddings are extracted from pre-trained neural networks. A fixed-length utterance-level representation is obtained from these embeddings. The fixed-length representation is finally classified by using an ensemble classifier. The utterance-level representation is obtained, either by computing the functionals that derive first order and second order moments, or by obtaining a bag-of-audio-word (BoAW) representation. The selection of the ensemble classification techniques was conducted similarly, as presented by one of the best performing techniques during Interspeech 2020 ComParE challenge (Markitantov et al., 2020). This enables us to compare systematically the neural embeddings against ComParE LLD representations.

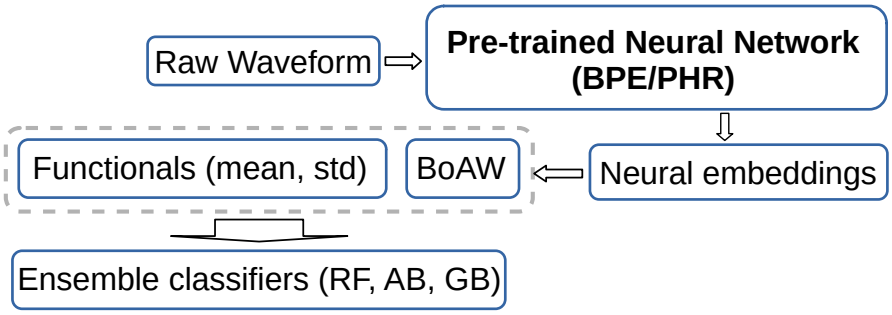


Figure 4.1 – The proposed neural embedding based method for COVID-19 detection. Mean and std denote the first order and second order moments used as functionals. RF denotes Random Forest (Ho, 1995), AB denotes Ada Boost (Freund et al., 1996) and GB denotes Gradient Boosting (Mason et al., 1999).

We investigate neural embeddings extracted from,

- Convolutional neural networks (CNNs) trained to model raw waveform for the task of phone classification in the context of speech recognition. One of the motivations behind using such an embedding is that COVID-19 infection affects speech production. As pointed earlier, in (Klumpp et al., 2021) it was found that there exist distinct phonetic patterns in COVID-19 infected speech. Although the authors conclude that those

patterns may not solely attribute to COVID-19 infection, it is still worth pursuing the idea.

- CNNs trained to model raw waveform for breathing pattern estimation. The main motivation behind that is that COVID-19 infection can adversely affect the functioning of respiratory system. As respiration process is intrinsic to speech production, breathing pattern information could be useful. In (Deshpande et al., 2021b), such idea was pursued with modeling of estimated breathing patterns from cough in an encoder-decoder framework. In this work, we do not model the output breathing patterns but rather we model the neural embeddings extracted from an intermediate layer.

4.2.2 Experimental setup

In this section, we first present the DiCOVA-II challenge dataset and the experimental protocols. Next, we present the extraction of different fixed-length representations, and finally the classifiers trained to detect COVID-19 infection.

4.2.2.1 Database and protocols

The data for the DiCOVA-II challenge (*Second DiCOVA challenge* 2021) is derived from the Coswara dataset (Sharma et al., 2020). Speech, cough, and breathing sound recordings from 956 subjects are organized in a 5-fold cross validation setting for development studies. Among these, 172 subjects were reported as tested positive for COVID-19 with mild to moderate symptoms or asymptomatic while the remaining 773 were reported as healthy with symptoms such as cold, cough, or fever, or with pre-existing respiratory conditions such as asthma. In addition, a blind test fold with 471 audio segments is provided to evaluate and report the performance of systems realized for the challenge.

The data is organized for four separate Tracks, following the same protocol. The data for the first Track includes 4.6 hours of breathing sound recordings. The second Track includes 1.7 hours of cough sound recordings, and the third Track includes 3.9 hours of speech recordings. There is no independent data for the fourth Track and fusions of the systems from either of the previous Tracks is considered for evaluating the performance over this Track.

The audio and speech signals in the dataset were resampled at 16 kHz, to derive features for our experiments. We used the pre-designed protocol across 5 folds to report our system on the development (*Dev*) set. We further accumulate the data across all folds for training to evaluate our system on the *Test* set.

4.2.2.2 Extraction of fixed-length feature representations

The feature representations used in this study are as follows:

- Interspeech 2013 ComParE LLDs based: The modeling of the LLDs was recently investigated as part of Interspeech 2021 ComParE sub-challenges for COVID-19 detection (Schuller et al., 2021). ComParE LLDs set with 65 frame-level features and their 65 delta coefficients (Δ), denoted as CMP_L, are extracted. Fixed-length representations are obtained by two methods: (a) by applying functionals over low-level acoustic descriptors (CMP_L) related to energy, spectral behaviour, and voicing based information, resulting in 6373 dimensional fixed-length feature vector denoted as CMP_F (refered to as ComParE feature set in Section 2.1.1.1) and (b) by extracting a BoAW representation of CMP_L with two separate codebooks size of 50; the first for static and the second for Δ LLDs, combined 100 dimensional representation denoted as $BoAW(CMP_L)$. The features are normalized to have zero median value. This normalization is done using the median and interquartile range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile). After this normalization, the statistical outliers are removed. We investigate these features for Track 1, Track 2 and Track 3.

- Embeddings from phone recognition neural network: we use an off-the-shelf CNN based neural network that models raw waveform to classify phones. This network takes 250ms of raw audio with a 10ms shift as an input, and consists of 10 convolutional layers followed by a hidden layer with 1024 nodes and an output layer. It is trained on AMI corpus (Carletta, 2007). Frame-level neural embeddings of 1024 dimensions, denoted as PHR, are extracted before activations of hidden layer and two different length representations are obtained (a) by computing functionals (mean and standard deviation) of the frame-level neural embeddings denoted as $f_{\mu\sigma}$(PHR), and (b) by extracting a BoAW representation of PHR with a codebook size of 100 denoted as $BoAW$(PHR). We investigate these features only on Track 2 and Track 3, as Track 1 only contains breathing recordings.

- Embeddings from breathing pattern estimation neural network: we use one of the raw waveform based CNNs trained in Section 3.3.2 that estimates breathing pattern at the output by taking 3 seconds of speech signal as input. This network consists of four convolution layers, one hidden layer with 10 nodes and one output unit. The network is trained on the Philips database (see Section 3.4.1) with mean squared error loss. Frame-level neural embeddings of 10 dimensions, denoted as BPE, are extracted before activation of hidden layer and two different length representations are obtained (a) by computing functionals (mean and standard deviation) of the frame-level neural embeddings denoted as $f_{\mu\sigma}$(BPE), and (b) by extracting a BoAW representation of BPE with a codebook size of 100 denoted as $BoAW$(BPE). We investigate these features only on Track 3 for the following reasons. Track 1 contains only breathing recording. Track 2 contains cough recordings. In a previous investigation carried out as part of the Interspeech 2021 ComParE challenge, we observed that the breathing pattern during cough is considerably different from the breathing pattern during speech production. Further investigation was needed to ascertain the utility of the information extracted.

OpenSMILE toolkit (Eyben et al., 2010) is used for extraction of CMP_L and CMP_F and openXBOW toolkit (Schmitt et al., 2017) for BoAW representation generation.

4.2.2.3 Classification

The ensemble classifiers used in this study are Random Forest (RF) (Ho, 1995), Ada Boost (AB) (Freund et al., 1996), and Gradient Boost (GB) (Mason et al., 1999) (see Section 2.1.2.1). We perform a grid search with AUC as optimization criterion to tune the hyperparameters of the classifiers using *Scikit-learn* (Pedregosa et al., 2011) toolkit. Tuning of hyperparameters for ensemble based classifiers is performed over the *Train* and *Dev* folds defined as per the challenge protocol. For most of the cases, RF classifier yielded the best performance. The AB classifier gave comparable yet lower performance for experiments on the *Dev* set, and hence the results in Section 4.2.3 are presented only for RF classifier. The parameters optimized for the RF classifier are the number of estimators {500, 1000, 2000}, maximal number of features {“auto”, “sqrt”, “log2”}, criterion {“gini”, “entropy”}, and minimal samples leaf {1, 2, 4}.

In addition to the framework with standalone features and classifiers, two fusion methods are also implemented to improve upon individual scores, namely, early fusion (EF) which is feature-level combination of fixed-length representations within a classification framework, and late fusion (LF), where the output probabilities from different systems are averaged before making a decision.

4.2.3 Results and analysis

Evaluation scores for our best performing systems for different tracks of the DiCOVA-II challenge are presented in Table 4.1. As per the challenge protocol, the metrics used for evaluation are the AUC and the sensitivity on the *Test* set in percentage (%). The reported sensitivity is obtained at a specificity of 95%. For each track, the results for the given baseline system in the challenge is also reported. The baseline classification system uses a bidirectional LSTM (BiLSTM) network to model log Mel spectrogram (*Second DiCOVA challenge* 2021).

Our system for Track 1, based on the ComParE features, performs comparable to the baseline across the *Dev* set. For the *Test* set, the best performing system is realized by a late fusion of RF scores obtained using two sets of ComParE features. Even though the AUC on the *Test* set is lower compared to the reported baseline system, our proposed system obtains considerably better sensitivity.

On Track 2 and Track 3, the ComParE feature based systems give lower performance when compared to PHR neural embeddings based systems. PHR embeddings yield the best performance. In terms of fixed-length representations, for PHR on Track 3, we observe that although both functionals and BoAW representations yield similar performance (also see Figure 4.3), BoAW representation yields better sensitivity. A late fusion of the scores obtained with the two representations marginally increases the AUC on the *Test* set, however does not contribute to

Table 4.1 – Results obtained for different systems over *Dev* and *Test* set of the DiCOVA-II challenge. The results are expressed in AUC metric and the sensitivity of the systems on the *Test* set at specificity of 95%. The systems noted as *[I]*, *[II]*, *[III]*, and *[IV]* were used in fusion method for Track 4.

System Feature	Classifier	*Dev* (%)	*Test* (%)	Sensitivity (%)
Track 1				
CMP_F	RF	77.83	76.78	30.0
$BoAW(CMP_L)$	RF	73.58	74.52	31.67
CMP_F , $BoAW(CMP_L)$ ***[I]***	LF	77.56	**78.05**	**43.33**
BASELINE	BLSTM	77.25	84.50	31.67
Track 2				
$BoAW$(PHR)	RF	70.06	74.19	30.0
$f_{\mu\sigma}$(PHR)	RF	70.54	72.87	26.67
CMP_L	RF	66.09	66.68	16.67
$f_{\mu\sigma}$(PHR), $BoAW$(PHR) ***[II]***	LF	71.32	**74.63**	**31.67**
BASELINE	BLSTM	75.21	74.89	36.67
Track 3				
$BoAW$(PHR) ***[III]***	RF	77.37	80.08	**41.67**
$f_{\mu\sigma}$(PHR)	RF	76.33	79.3	26.67
$BoAW$(BPE)	RF	68.93	73.49	21.67
$f_{\mu\sigma}$(BPE)	RF	68.44	—	—
$BoAW(CMP_L)$	RF	70.38	75.59	15.0
EF($f_{\mu\sigma}$(PHR), $f_{\mu\sigma}$(BPE)) ***[IV]***	RF	76.67	79.1	28.33
EF($BoAW$(PHR), $BoAW$(BPE) , $BoAW(CMP_L)$)	RF	77.47	79.95	33.33
$f_{\mu\sigma}$(PHR), $BoAW$(PHR)	LF	77.59	**80.64**	36.67
BASELINE	BLSTM	80.16	84.26	43.33
Track 4				
[III], [IV]	LF	77.79	**80.51**	40.0
[I], [IV]	LF	80.09	78.05	**43.33**
[I], [II]	LF	77.93	78.05	**43.33**
BASELINE	LF	81.67	84.70	55.0

the sensitivity. Similar observations can be noted for Track 2 except that late fusion of these representations slightly increases both AUC and sensitivity on the *Test* set.

Looking into the system performances for Track 4, it appears that the classifiers trained using breathing sounds are more prominent when fused with systems trained with cough and speech signals. The fused systems in Track 4 has higher AUC compared to the best system from Track 2 and comparable AUC to the best system from Track 3.

4.2.3.1 Analysis of neural embedding based systems

Figure 4.2 shows the cumulative frequency response of the first convolutional layer for PHR CNN and BPE CNN. It can be observed that the PHR network emphasizes around the formant frequency regions in speech, while the emphasis of the BPE network is significantly towards the lower frequency region. In other words, they are modeling different information from the speech signal.

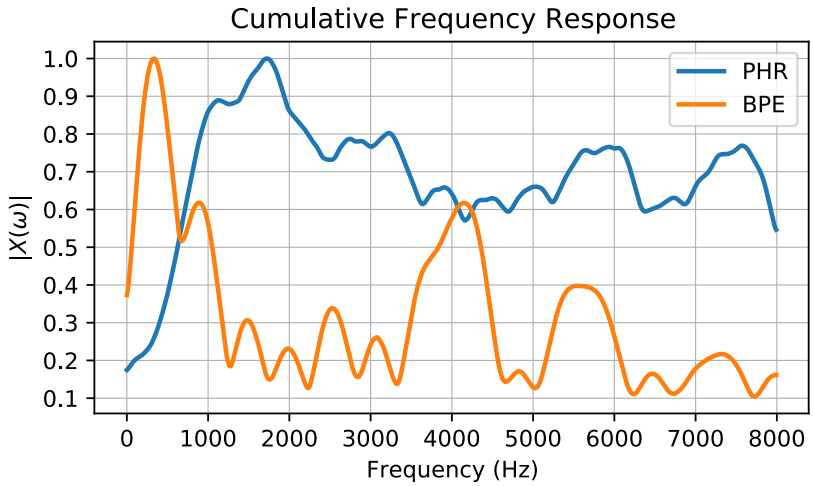


Figure 4.2 – The cumulative frequency response of the kernels for the first convolution layer of the CNN models pre-trained for phone recognition (PHR) and breathing pattern estimation (BPE).

Figure 4.3 shows the ROC plot for both feature sets of PHR and BPE based networks for Track 3. It can be seen that the PHR neural embeddings consistently yield better system than BPE neural embeddings. One of the reason for that could be that the participants may not have had severe COVID issues for the differences with respect to non-COVID participants to be very apparent at BPE embedding level.

4.2.3.2 Analysis of LLD based systems

In order to analyse the discriminability of the features used for our studies, we estimated their respective importance in achieving the desired classification performance. This analysis improves our comprehension towards the significance of certain features for identifying COVID-19 positive cases in a given audio modality. Table 4.2 presents an overview of the most important features for each track, based on an importance scores generated by the RF based classifier.

For all the tracks, the auditory spectra coefficients obtained using RASTA filtering (audSpec:Rfilt) and their deltas (Δ) establish as one of the most discriminative LLDs. For Track 1, coefficients obtained as the third quartile of these features prove significant for classification. For Track 2, an extended list of functionals prove significant with features capturing primarily

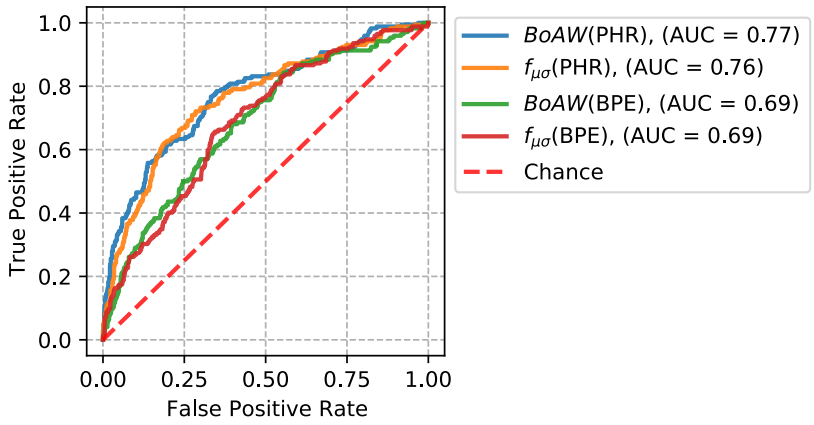


Figure 4.3 – ROC plot for systems trained using PHR embeddings and BPE embeddings on the *Dev* set of Track 3.

the spectral shape. For Track 3, speech specific features such as MFCC and spectral band energy prove discriminatory for the task.

Table 4.2 – LLDs and functionals exhibiting highest discriminability for each track (most representative, non-redundant features).

LLDs	**functional**
Track 1	
Δ audSpec_Rfilt	3^{rd} quartile
voicing parameters	LP–gain
magnitude spectra	RollOff
Δ magnitude spectra	variance
Track 2	
audSpec_Rfilt	regression coefficients, centroid, 2^{nd} quartile
Δ Pitch contour	regression coefficients
Δ RMSenergy	extremums
band energy magnitude spectra	extremums
magnitude spectral slope	regression coefficients
Track 3	
audSpec_Rfilt	regression coefficients, 1^{st} quartile
mfcc	peak behavior , percentiles
Δ audSpec_Rfilt	peak behavior
Δ magnitude spectra	moments

4.2.4 Summary of the study

In this study, we investigated modeling of neural embeddings extracted from raw waveform based neural networks, pre-trained for phone classification and breathing pattern estimation,

for the task of COVID-19 infection detection. More precisely, these embeddings were modeled as fixed-length representations through application of functionals and BoAW, similar to modeling of knowledge-based LLDs for paralinguistic speech processing. Our investigations on the DiCOVA-II challenge demonstrated that neural embeddings derived from the phone classification neural network (PHR) outperformed systems based on knowledge-based LLDs and those using breathing pattern estimation (BPE) embeddings. In Track 3, we observed that while BPE embedding based systems achieved slightly lower overall performance compared to LLD based systems, they exhibited better sensitivity. Overall, our findings suggest that modeling neural embeddings from networks trained on auxiliary or related speech tasks is a promising approach for COVID-19 infection detection and has the potential to replace traditional knowledge-based features.

4.3 Breathing Pattern in Synthetic Speech

Besides natural speech produced by humans, advances in speech technology also have made it possible to generate speech. Text-to-speech synthesis (TTS) methods have evolved over time. For a long time concatenative TTS (Taylor, 2009) and statistical parametric TTS (Zen et al., 2009) were the main methods of generating speech from text but recently there has been a shift towards deep learning based methods (Oord et al., 2016; Y. Wang et al., 2017). The speech generated by deep learning based methods are reportedly very natural sounding and in some cases are indistinguishable from human speech.

A research question that arises is: whether synthetic speech carries breathing related information in the same way as natural human speech? Besides scientific curiosity, answer to this research question is of potential interest to other complementary research directions, such as, (a) TTS systems can be used to fake identity, e.g. presentation attack on automatic speaker verification systems (Z. Wu et al., 2015) and (b) TTS is being explored for synthesizing speech with pathological conditions to develop objective pathological speech methods (Halpern et al., 2021). In pathological speech such as dysarthric speech, breathing phenomenon is intrinsically related to impaired speech production (Duffy, 2013; Enderby, 1980). This study aims to address the aforementioned research question by leveraging two different research directions, namely, speech based breathing pattern estimation and detection of logical access presentation attack.

4.3.1 Study design

In this study, we investigate the aforementioned research question using pre-trained raw waveform CNNs for estimating breathing pattern estimation that we developed in Chapter 3. However, one issue is that to evaluate the output breathing pattern we need a reference breathing pattern to compare to, which in the case of human speech production can be measured through sensors but not in the case of synthetic speech. In Section 4.2, we demonstrated that neural embeddings extracted from pre-trained breathing pattern estimation neural networks

could be used for COVID-19 detection. Based on this observation, we recast the research question as: whether natural human speech and synthetic speech can be distinguished based on embeddings extracted from pre-trained breathing pattern estimation neural networks. The underlying hypothesis being that: should synthetic speech exhibit breathing pattern information similar to natural human speech then the two speech signals will not be easily distinguishable.

Figure 4.4 illustrates our framework. The embeddings extracted from a pre-trained neural network for breathing pattern estimation (BPE) are used to train a binary classifier, natural versus synthetic speech. We investigate two classification approaches. In the first approach, frame-level neural embeddings are classified using a multi-layer perceptron (MLP). The output class probabilities are averaged over the utterance and a decision is made. In the second approach, similar to the aggregation method proposed in Section 4.2.1, the embeddings are aggregated using functionals (mean, standard deviation) or bag-of-audio-words (BoAW) representation to obtain an utterance-level fixed-length representation and then classified using classifiers such as, support vector machine (SVM) and Random Forest (RF).

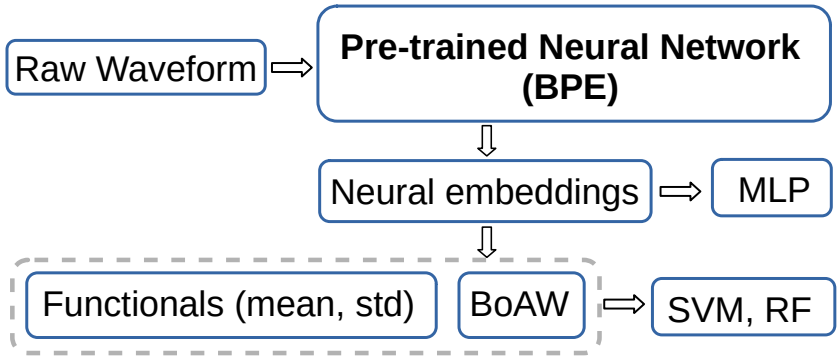


Figure 4.4 – Framework to distinguish natural human speech and synthetic speech based on breathing pattern embeddings.

To investigate this question, we leverage from the automatic speech verification community's effort in developing anti-spoofing methods to detect logical access attacks generated using TTS systems and voice conversion (VC) systems through organization of ASVspoof challenge (*ASVspoof* 2022; Todisco et al., 2019). Besides well-defined protocols and employing state-of-the-art approaches to generate logical access attacks, the ASVspoof challenge provides the means to systematically investigate the research questions. First, there are two types of TTS systems (Todisco et al., 2019): (a) synthetic speech purely generated using neural models trained on speech and textual data (neural TTS) and (b) synthetic speech generated by concatenating segments of natural human speech waveforms (concatenative synthesis). Second, there are also attacks generated through voice conversion alone (VC-alone). The VC-alone system takes a natural human speech signal as input and converts it to the target speaker's voice by altering the source and system information. So, as a by-product, it allows us to investigate whether such alterations done on a single speaker speech affect the breathing pattern related information.

4.3.2 Experimental setup

This section first presents the ASVspoof 2019 database and protocol. Next presents extraction of neural embeddings using pre-trained breathing pattern estimation neural networks, and finally the development of binary classifiers (natural human speech vs. synthetic speech).

4.3.2.1 Database and protocols

We use the ASVspoof2019 challenge (Todisco et al., 2019) database for our investigation. The ASVspoof2019 challenge provides presentation attacks for two use case scenarios: logical access (LA) and physical access (PA). Our investigation focused on the LA scenario in which attacks are generated using TTS and VC technologies. The database includes three sets, namely *Train*, *Dev*, and *Test* which comprise of speech from 20 (8 male, 12 female), 10 (4 male, 6 female) and 48 (21 male, 27 female) speakers respectively. The *Train* set includes 2580 bonafide and 22800 spoofed utterances while the *Dev* set comprises 2548 bonafide and 22296 presentation attacks. The *Test* set includes 7355 bonafide and 63882 presentation attacks.

The database includes bonafide and presentation attacks generated by 17 different TTS and VC systems. From these systems, 6 are considered as known attacks which are the only presentation attacks present in the *Train* and *Dev* sets. The remaining 11 are considered unknown attacks. The 11 unknown attacks and 2 of the known attacks comprise presentation attacks available in the *Test* set. The VC systems use neural network based and spectral filtering based approaches (Matrouf et al., 2006). Concatenation based and neural network based systems are used for TTS systems with different vocoders (Oord et al., 2016; Morise et al., 2016). The TTS-VC systems use various waveform generation methods such as Griffin-Lim (Griffin et al., 1984) and generative adversarial networks (Tanaka et al., 2018) among others.

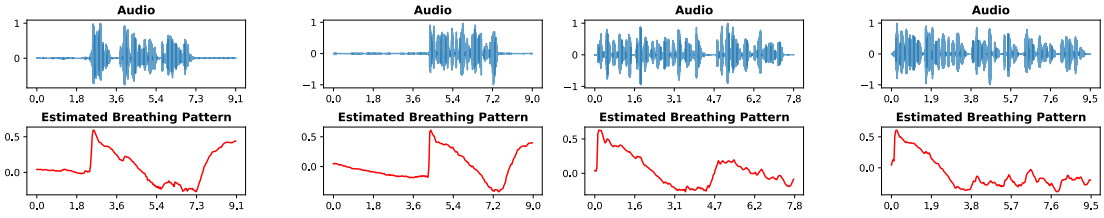
In all our experiments, we follow the protocols as per ASVspoof2019 challenge, i.e., the *Train* and *Dev* sets are used for training the binary classifiers and evaluation is carried out on the *Test* set. We use equal error rate (EER) and area under the receiver operating characteristic curve (AUC) as the evaluation measures.

4.3.2.2 Neural embeddings based feature representation

For the present study, we use the raw waveform CNNs trained for breathing pattern estimation that we developed in Chapter 3. In Chapter 3, we reported the best performing CNN models, however, we observed that the performance of the CNNs in breathing pattern estimation were comparable. So, for the present study, we chose two CNNs each trained on Philips database and UCL-SBM database with mean squared error loss function, one with 2 seconds speech as input and the other with 3 seconds speech as input. The model trained on the Philips database with 3 seconds speech as input was used in Section 4.2 for COVID-19 detection. Figure 4.5 shows the estimated breathing pattern output by the 3 seconds input CNN pre-trained on the Philips database for a bonafide speech, VC attack, TTS attack and TTS_VC attack from the

ASVspoof2019 database. When extracting the 10-dimensional neural embeddings from the hidden layer (before activations), for the utterances with duration of shorter than twice the size of the input window (i.e., 2 seconds or 3 seconds), we simply repeat the whole utterance and include them in our study to avoid changing the ASVspoof2019 challenge protocol. This is acceptable as we are not interested in extraction of breathing patterns in an absolute sense.

As mentioned earlier, the neural embeddings are modeled by different classification techniques. In the case of the MLP classifier, no further processing is needed. In the case of fixed-length representation we follow the same procedure as in Section 4.2.2.2. We compute the utterance-level mean and standard deviation as functionals and concatenate them obtain a 20-dimensional representation denoted as $f_{\mu\sigma}$(BPE). In the case of BoAW based fixed-length representation, denoted as $BoAW$(BPE), we use the openXBOW toolkit (Schmitt et al., 2017) with a codebook size of 100 to obtain 100 dimensional BoAW representation.



(a) – Bonafide **(b)** – VC **(c)** – TTS **(d)** – TTS_VC

Figure 4.5 – Estimated breathing pattern with a CNN pre-trained on Philips database with input speech window length of 3 seconds for different examples from ASVspoof2019 database with natural and synthetic speech.

4.3.2.3 Classification framework

Three different classifiers are trained with the features obtained from the CNNs as explained in 4.3.2.2. The 10-dimensional frame-level embeddings denoted as BPE are classified using an MLP with two fully connected layers. The input layer consists of 10 nodes. The first and second hidden layers consists of 128 and 64 nodes, respectively. The MLP is trained using the binary cross entropy loss function and the Adam optimizer (Kingma et al., 2015) with a learning rate of 0.001. The system is implemented using PyTorch (Paszke et al., 2019) framework.

The utterance-level embeddings denoted as $f_{\mu\sigma}$(BPE) and $BoAW$(BPE) are modeled by an SVM with linear kernel and a Random Forest (RF) classifier. Similar to classification framework in Section 4.2.2.3 we fine-tune the hyperparameters of the classifiers using grid search methodology with AUC as optimization criterion. The SVM is tuned for different values of the regularization parameter C. For RF classifier, the parameters for grid search are as following: number of estimators {50, 500, 1000, 2000}, maximal number of features {“auto”, “sqrt”, “log2”}, criterion {“gini”, “entropy”}, and minimal samples leaf {1, 2, 4}. In all cases we used the *StandardScalar* method of *Scikit-learn* for normalizing the data.

Table 4.3 – The AUC and EER in percentage on the evaluation set for embeddings obtained from CNNs pre-trained on the Philips database with input speech window length of 3 seconds and 2 seconds. Column "All" presents the system performance over all the evaluation data while the results under other columns are reported over a subset of evaluation data with all the bonafide files and only the presentation attacks with the type mentioned as the title of the column. VC stands for voice conversion, TTS for Text-to-speech, and TTS_VC is a combination of the two.

Features	Classifier	Measure	All	VC	TTS	TTS_VC
Embeddings from CNN pre-trained on Philips database						
3 seconds speech input						
BPE	MLP	AUC	90.35	59.92	99.4	99.7
		EER	16.88	42.75	2.53	1.32
$f_{\mu\sigma}$(BPE)	SVM	AUC	89.51	59.42	98.42	98.78
		EER	16.98	43.54	4.29	3.48
	RF	AUC	90.65	62.44	98.93	99.54
		EER	17.02	41.02	4.22	2.6
$BoAW$(BPE)	SVM	AUC	89.35	61.43	97.54	98.17
		EER	17.69	41.62	7	6.01
	RF	AUC	90.86	62.72	99.16	99.62
		EER	17.69	40.04	4.14	2.5
2 seconds speech input						
BPE	MLP	AUC	84.56	47.94	95.08	96.63
		EER	21.5	51.59	10.89	8.72
$f_{\mu\sigma}$(BPE)	SVM	AUC	87.52	57.92	95.85	97.7
		EER	20.08	44.39	10.63	7.49
	RF	AUC	89.15	56.68	98.61	99.55
		EER	18.23	45.41	5.25	2.7
$BoAW$(BPE)	SVM	AUC	88.18	52.17	98.91	99.18
		EER	19.28	48.57	4.24	2.97
	RF	AUC	88.04	51.14	99.01	99.34
		EER	19.51	48.46	4.61	3.2

4.3.3 Results and analysis

Tables 4.3 and 4.4 shows the AUC and EER in percentage on the evaluation set for the features obtained from CNNs pre-trained on both Philips and UCL-SBM databases. The system performance over all the samples in the evaluation set is presented under the column "All". The results under other columns are reported over a subset of the evaluation data with all the bonafide files and only specific types of presentation attacks, namely VC, TTS, and TTS_VC.

When the whole evaluation set is taken into consideration (i.e., "All" column), it can be observed that irrespective of the database on which the CNNs are trained, input speech length window or type of classifier, the AUC ranges between 84.56% and 90.86% and the EER ranges between 16.88 % and 21.5 %. Looking into the results segregated in terms of the attack types reveals that TTS and TTS_VC can be classified relatively easier based upon BPE than VC. We achieve AUCs as high as 99.64% and 99.93% with EERs as low as 1.6% and 0.59% for TTS and

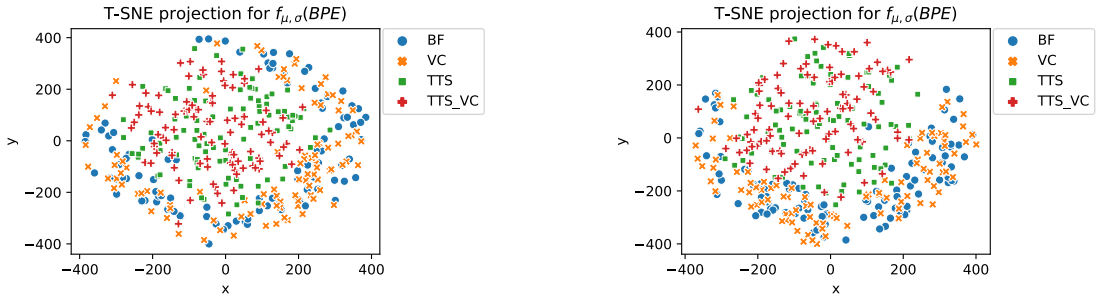
TTS_VC attacks, respectively, while a best AUC of 58.32% and EER of 44.44% is achieved for VC attack. The performance for VC is close to the chance level.

Table 4.4 – The AUC and EER in percentage on the evaluation set for embeddings obtained from CNNs pre-trained on the UCL-SBM database with input speech window length of 3 seconds and 2 seconds. Column "All" presents the system performance over all the evaluation data while the results under other columns are reported over a subset of evaluation data with all the bonafide files and only the presentation attacks with the type mentioned as the title of the column. VC stands for voice conversion, TTS for Text-to-speech, and TTS_VC is a combination of the two.

Features	Classifier	Measure	All	VC	TTS	TTS_VC
Embeddings from CNN pre-trained on UCL-SBM database						
3 seconds speech input						
BPE	MLP	AUC	90.02	58.33	99.43	99.73
		EER	17.27	43.65	2.56	1.65
$f_{\mu\sigma}$(BPE)	SVM	AUC	90.23	59.86	99.2	99.66
		EER	17.48	42.96	3.62	2.35
	RF	AUC	90.76	60.85	99.64	99.93
		EER	17.29	41.77	1.6	0.59
$BoAW$(BPE)	SVM	AUC	90.11	58.32	99.58	99.8
		EER	17.29	44.44	2.75	1.74
	RF	AUC	90.5	60.11	99.49	99.9
		EER	17.07	42.62	2.46	0.63
2 seconds speech input						
BPE	MLP	AUC	89.84	60.24	98.42	99.41
		EER	17.48	43.14	5.56	3.08
$f_{\mu\sigma}$(BPE)	SVM	AUC	88.07	58.28	97.25	96.45
		EER	18.97	43.92	8.67	9.44
	RF	AUC	88.48	54.97	98.55	98.46
		EER	18.71	47.1	5.85	6.25
$BoAW$(BPE)	SVM	AUC	89.25	55.91	99.21	99.34
		EER	17.69	45.79	3.53	3.03
	RF	AUC	90.01	57.92	99.62	99.7
		EER	17.19	43.72	2.49	2.44

We observe the same trend whether we classify a single BPE frame through an MLP and aggregate the output probabilities or first aggregate the BPE into an utterance-level fixed-length representation through computation of first order and second order statistics or BoAW representation and then classify with SVM or RF. So, to get an insight into the BPE space, we generated a T-SNE (Maaten et al., 2008) projection visualization for $f_{\mu\sigma}$(BPE) features extracted from CNNs with 3 seconds speech input pre-trained on Philips and UCL-SBM databases. Figure 4.6 presents the T-SNE projection. The plots are generated for the $f_{\mu\sigma}$(BPE) features extracted from CNNs pre-trained on both Philips and UCL-SBM databases.

It can be observed that in both cases the samples from bonafide and VC are grouped together (blue circles and orange crosses) and TTS and TTS_VC are grouped together (green squares and red pluses). Furthermore, even though the distribution is different between the embeddings



(a) – Philips **(b)** – UCL-SBM

Figure 4.6 – TSNE projection of $f_{\mu\sigma}$(BPE) embeddings extracted from CNNs pre-trained on (a) Philips and (b) UCL-SBM database with 3 seconds input speech.

extracted from the two CNNs, there is a good separation between the two groups. It is worth mentioning that the T-SNE projections with 2 seconds input CNNs also yielded the same observations. The T-SNE visualizations reaffirm the observations made from the results presented in the Tables 4.3 and 4.4.

Table 4.5 – The median values for the AUC and EER for our systems and the median values for the EER of the systems presented in ASVspoof2019 challenge. The values are presented in percentage. The numbers in brackets are the range of EER for our systems.

	Attack type	Our study		ASVspoof2019
		AUC	EER	EER (Todisco et al., 2019)
A07	TTS	99.75	1.48 [0.38 - 5.39]	0.02
A08	TTS	98.65	4.76 [2.38 - 7.55]	0.09
A09	TTS	99.3	3.31 [0.45 - 12.36]	0.06
A10	TTS	99.74	1.39 [0.74 - 5.44]	12.21
A11	TTS	99.68	1.64 [0.79 - 5.35]	0.59
A12	TTS	98.85	4.6 [1.6 - 10.13]	3.75
A13	TTS_VC	99.75	1.31 [0.31 - 7.22]	12.41
A14	TTS_VC	99.67	2.13 [0.73 - 4.82]	2.88
A15	TTS_VC	99.68	1.79 [0.53 - 4.78]	3.22
A16	TTS	99.32	3.27 [1.69 - 5.37]	0.02
A17	VC	52.96	48.07 [42.93 - 51.22]	15.93
A18	VC	61.82	41.51 [38.9 - 45.38]	5.59
A19	VC	65.8	38.25 [35.66 - 39.74]	0.06

We further analyse the AUC and EER per attack on the *Test* set by computing median value for AUC and EER obtained with both Philips and UCL-SBM CNNs trained with 3 seconds speech input. Table 4.5 presents median values along with the range of EER and AUC per attack and contrasts with the median EER reported in the ASVspoof2019 challenge for all the systems in (Todisco et al., 2019). It can be observed that the systems in ASVspoof2019 challenge have yielded high EER on attacks A10 (TTS), A13 (TTS_VC) and A17 (VC). In our case, high EERs are only observed for VC attacks. It can be also noted that on some attacks, namely, A10,

A13, A14 and A15, the median EERs in our study are lower than the median EER obtained in ASVspoof2019 challenge.

4.3.4 Summary of the study

In this study, we investigated whether synthetic speech carry breathing pattern related information in the same way as natural human speech. We investigated this question by conducting a study on ASVspoof2019 challenge to distinguish between bonafide speech and attacks generated through TTS, combination of TTS and VC, and VC using breathing pattern embeddings estimated using networks pre-trained on two different databases, one with read speech and one with conversational speech. Our results and analyses consistently showed that attacks based on TTS speech and TTS_VC speech can be detected in a highly accurate manner when compared to attacks based on VC-alone. This indicates that, irrespective of the TTS approach i.e. whether concatenative synthesis or neural TTS, the generated synthetic speech tends to not carry breathing pattern related information in the same way as natural human speech. Furthermore, the findings also indicate that the alterations done to the natural human speech signal during voice conversion is not strongly altering the breathing pattern related information.

4.4 Emotion Recognition

Emotions are quintessential elements of communication among humans, and are expressed in different ways across several modalities. Speech is one of the prime modes to convey the expression of emotions, and hence emotion recognition using the acoustic content of signal is gaining popularity in speech application areas. Human emotions are paralinguistic phenomena which manifest distinctively over varying temporal and spectral characteristics. Due to limitations with representation and processing, extraction of human emotions using traditional acoustic signal analysis method is challenging. Recent trends have witnessed a growing interest in the field of multi-modal emotion recognition, which attracts extensive use of deep neural networks (DNNs) to exploit the contrast between speech, textual, and physiological modalities (Tzirakis et al., 2017; Tzirakis et al., 2021; Schuller, 2018; Stappen et al., 2020; J. Zhang et al., 2020; Shen et al., 2020; Khare et al., 2020).

Physiological signals have been used for emotion recognition (Dhall et al., 2020). Within the field of affective computing, recognition approaches to predict continuous states of emotion, frequently utilize the two-dimensional Circumplex Emotion Model (Russell, 1980), observing the valence and arousal of speaker's emotional state. However, in order to avoid subjective emotional labels, multiple raters must continuously annotate, which is costly and time-expensive. In (Baird, Stappen, et al., 2021), authors showed that bio-signals processed with MuSe-Toolbox (Stappen, Schumann, et al., 2021) could be used as alternative to Evaluator Weighted Estimator (EWE) (Grimm et al., 2005) emotional gold standard (Baird et al., 2019). Experimental study presented by (Boiten et al., 1994) shows that respiration patterns reflect

the general dimensions of emotional response.

This study has been done as part of the Multimodal Sentiment Analysis (MuSe) 2022 challenge. It focuses on the MuSe-Stress sub-challenge, where the goal is to predict emotion in a time continuous manner. In that context, we investigate, modeling of embeddings extracted from the breathing pattern estimation (BPE) networks for emotion recognition. Here we present the performance of these embeddings against the embeddings extracted from networks trained for phone recognition (PHR) and speech emotion recognition (SER) tasks. This challenge provided physiological signals along with the audio signal. We also present the performance of an end-to-end CNN based system modeling physiological signals to estimate valence and arousal as well as the fusion of the systems. The complete investigation is presented in (Yadav et al., 2022). My contribution in the study has been on the use of embeddings extracted from breathing pattern estimation networks. The other results presented here are for comparison. For complete results, see (Yadav et al., 2022).

4.4.1 Proposed approaches

We pursue two approaches, namely, (a) modeling embeddings extracted from acoustic signal using pre-trained neural networks (Figure 4.7) and (b) an end-to-end CNN based system modeling physiological signals (Figure 4.8) to estimate valence and arousal. Furthermore, we investigate fusion of acoustic information and physiological information.

4.4.1.1 Pre-trained feature representations

Figure 4.7 illustrates our method. In this approach, similar to the frameworks in Section 4.2.1 and Section 4.3.1, first, frame-level neural embeddings are extracted from pre-trained networks using the acoustic signal. Fixed-length representation is then obtained for each 500 ms of signal by applying functionals (mean and standard deviation). Arousal and valence are then estimated as a regression problem by feeding the fixed-length representation as input to a neural network.

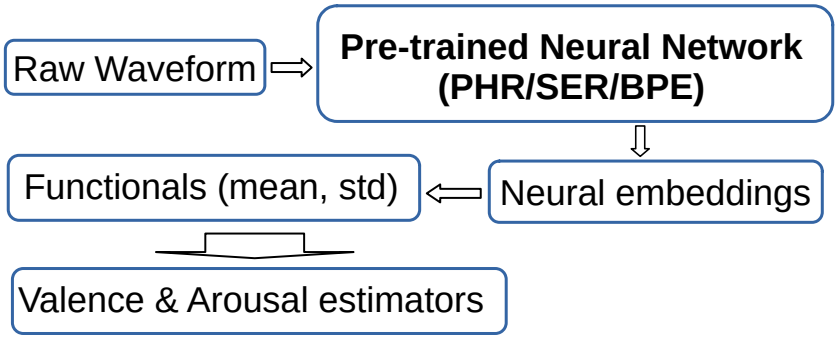


Figure 4.7 – Proposed pipeline for using embeddings from pre-trained networks.

We investigate neural embeddings extracted from:

- Convolutional neural networks (CNNs) trained for phone classification from raw waveform speech. Phonetic information has been shown to capture emotional content in speech. Previously, (Vlasenko et al., 2011) showed a strong correlation between the vowel formants and the level of arousal in human speech, while (Shah et al., 2019) demonstrated that incorporating speech articulatory information improves valence based classification. The benefit of modelling speech and phonetic units for speech emotion recognition (SER) task was shown in (Schuller et al., 2008; Yuan et al., 2021). All these prior works inspired us to use phone based embedding for the task-at-hand.
- Convolutional neural networks (CNNs) trained for speech emotion recognition (SER) task. Since the task-at-hand deals with predicting two of the emotion dimensions, emotional valence and arousal, it deemed appropriate to generate embeddings from a network trained for SER task.
- Convolutional neural networks (CNNs) trained for breathing pattern estimation task. Speech carries a wide range of information including age (Bocklet et al., 2008), gender (Vergin et al., 1996), and emotional state (El Ayadi et al., 2011) of a person. Respiration is one of the physiological signals altered by emotion. Relationships between emotions and respiratory patterns have shown more rapid breathing during an speaker's emotional arousal state (Boiten, 1998). Respiration with deep learning based methods has been used for emotion recognition (Q. Zhang et al., 2017). There is a close relation between speech and breathing as well since speech is produced by organs evolved for breathing (MacLarnon et al., 1999). The embeddings extracted from the CNN models pre-trained for estimating breathing patterns have shown to be informative for auxiliary tasks such as detection of COVID-19 (Section 4.2) and distinguishing between natural and synthetic speech (Section 4.3). Based on these observations we hypothesize that such embeddings can be used for emotion recognition.

4.4.1.2 Arousal and valence estimator

The sequential nature of the selected regression tasks makes recurrent neural networks (RNNs) a natural choice for a comparably simple system. We use the LSTM-RNN system proposed as part of baseline system provided by the challenge organizers without any modification to the architecture or training process for estimating valence and arousal. As done in the baseline studies, we train a separate estimator for valence and arousal. It allows us to fairly compare our proposed embeddings and feature representations to those of the baseline studies.

4.4.1.3 Modelling raw physiological signals using CNNs

Several works propose modelling raw waveform signals for various tasks, such as speech recognition (Palaz et al., 2013; Sainath et al., 2015; Collobert et al., 2016), speaker recognition (Muckenhirn et al., 2018; Ravanelli et al., 2018), gender recognition (Kabil et al., 2018), depression detection (Dubagunta et al., 2019), and audio classification (Zeghidour et al., 2021).

We use a raw waveform based approach for modeling physiological signals in Chapter 3. In the same light, we propose a CNN based framework for directly modelling raw physiological signals for estimating valence and arousal in an end-to-end manner. The physiological signals include ECG, beats per minute (BPM), and respiratory signal (RESP). Figure 4.8 illustrates this method. Given the nature of these input signals as well as the annotation granularity of the data (every 500 ms), each physiological signal is modelled after centering of the labelled segment with appropriate context.

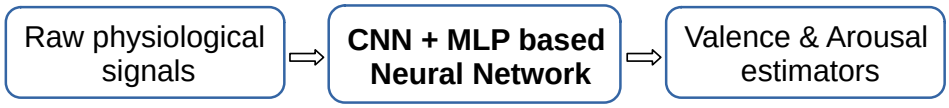


Figure 4.8 – Proposed pipeline for end-to-end system to estimate valence and arousal from raw physiological signals.

4.4.1.4 Fusion based estimation

We also investigate a combination of different embeddings extracted from acoustic signal and physiological signals. We investigate early fusion, where different features are concatenated and are fed as input to the RNN based arousal and valence estimator, presented earlier in Section 4.4.1.2.

4.4.2 Experimental setup

This section describes the experimental setup for the proposed study, including a brief description of the dataset and the evaluation protocol and the evaluation metric used for the study. This is followed by a description of the training methodology and ablation experiments for development of the proposed raw physiological signal modelling CNNs.

4.4.2.1 Dataset and protocol

The MuSe-Stress sub-challenge is a regression task on continuous signals for emotional arousal and valence (Christ et al., 2022). The Ulm-Trier Social Stress Test dataset (Ulm-TSST) is used to set up training, development, and testing subsets, comprising individuals in stressful situations following the Trier Social Stress Test (TSST) (Kirschbaum et al., 1993). The dataset provides *Train*, *Dev* and *Test* splits with 41, 14 and 14 subjects, respectively. We further split the training set into *training* set with 32 subjects and *validation* set with 9 subjects for tuning hyperparameters.

4.4.2.2 Evaluation metrics

The regression task MuSe-Stress is evaluated in terms of concordance correlation coefficients (CCC) (see Section 2.3.2) for arousal, valence and combined modalities. The ultimate goal of

the challenge is to reach the highest possible combined *CCC* score.

4.4.2.3 Baseline systems

The challenge organizers provided systems which can be used to evaluate the classification performance of features and networks for emotion dimensions. Also, extracted feature representation for audio (DEEPSPECTRUM (Amiriparian et al., 2017)) and bio-signal (ECG, BPM, and RESP) were provided by organizers.

4.4.2.4 Extracting fixed-length feature representations

As we mentioned earlier, we investigate embeddings extracted from three different pre-trained networks:

- Phone recognition neural network: We use the same network as presented in Section 4.2.2.2 for phone recognition. The 2048-dimensional vector obtained from computing functionals (mean and standard deviation) over frame-level embeddings is denoted as RAW(PHN).
- Speech emotion recognition neural network: For this we resort to an off-the-shelf CNN network similar to the raw waveform CNN network presented in (Purohit, Yadav, et al., 2023) that models raw audio signals in an end-to-end manner. The network is trained for SER task using IEMOCAP corpus (Busso et al., 2008) and consists of 4 convolutional layers followed by a fully connected layer with 10 nodes and an output layer with softmax activation for a 4 class classification corresponding to 4 emotion categories namely sad, happy, angry, and neutral. The input to the system is 250 ms of raw audio signal. The fixed-length representation after applying functionals, denoted as RAW(SER) is a 20-dimensional vector.
- Breathing pattern estimation neural network: We use one of the networks trained in Section 3.3.2 for estimating breathing pattern from speech signal. The network is trained on UCL-SBM database (see Section 3.4.1) and takes 3 seconds of speech signal as input. The embeddings are extracted before the activation of the fully connected layer for every 40 ms. The fixed-length representation is obtained for every 500 ms by applying functionals. The resulting 20-dimensional vector is denoted as UCLBS.

4.4.2.5 Training raw physiological signal CNNs

Table 4.6 depicts the general CNN architecture used for modelling respiratory (RESP-CNN), BPM (BPM-CNN) and ECG (ECG-CNN) physiological signals. The proposed CNN architecture consists of 4 convolutional layers followed by an MLP with one hidden layer. All the hidden layers are followed by a ReLU activation function. The number of filters in each layer as well

Table 4.6 – CNN architectures for physiological signals. Convolution parameters are denoted as Conv(filters, kernel width, stride), and MP denotes max-pooling layer. FC denotes fully connected layer.

RESP-CNN	BPM-CNN	ECG-CNN
Conv(64, 75, 15)	Conv(16, 175, 10)	Conv(56, 100, 15)
MP(2, 2)	MP(2, 2)	
Conv(128, 10, 1)	Conv(32, 10, 1)	Conv(112, 10, 1)
MP(2, 2)	MP(2, 2)	
Conv(256,7,1)	Conv(64,7,1)	Conv(224,7,1)
Conv(512,7,1)	Conv(128,7,1)	Conv(448,7,1)
FC(75)	FC(175)	FC(100)

as the kernel and stride parameters of the first convolution layer, which are dependent on the signal characteristics, are tuned individually for each physiological signal. A Dropout layer (Srivastava et al., 2014) is added before the MLP for improved regularisation.

For the challenge, the provided raw physiological signals had a sampling frequency of 1000 Hz, and labels were provided for every 500 ms intervals. Each physiological signal is centered with a *context* such that the center 500 ms segment of the input signal corresponds to the target label, normalized and is then directly fed to the model. We adopt a multi-task learning paradigm for optimizing physio-arousal and valence simultaneously, i.e., the final layer of each CNN returns 2 outputs, and the CNN is trained by optimizing the average combined CCC. Each CNN is trained in a multi-task setting with an AdamW optimizer (Loshchilov et al., 2019), with early stopping.

For each physiological signal CNN, the best number of filters in each layer as well as the dimension of the fully connected (FC) layer is tuned individually. We also optimize the context size that is best performing for each physiological signal CNN. Respiratory signal works best with the largest context of 2500 ms, which is inline with observations made in Chapter 3. The best context size for ECG-CNN and BPM-CNN is 500 ms and 1000 ms, respectively. More details on training the physiological signal CNNs can be found in (Yadav et al., 2022).

4.4.3 Results and analysis

This section describes the results obtained using various uni- and multi-modal feature representations. Table 4.7 shows the results of the various systems on the MuSe-Stress *Dev* and *Test* set. The top rows depict the best baseline result (DEEPSPECTRUM) and the best physiological signal based system as per the challenge paper (Christ et al., 2022), followed by our results using features obtained from the proposed physiological CNNs and acoustic representations. Finally, select multi-modal early fusion results are provided. All methods described are built on top of extracted features and their combinations using the provided baseline code.

Table 4.7 – *CCC* scores obtained on the *Dev* and the *Test* set by various systems. Best scores over 5 random seeds reported, with (mean ± std) over runs for the *Dev* set. "+" denotes early fusion, i.e. concatenation of the denoted features, respectively. *ndims* denotes feature dimensionality. The highset scores in each category are highlighted in bold.

		Arousal [CCC]		Valence [CCC]		Combined [CCC]
Features	ndims	*Dev*	*Test*	*Dev*	*Test*	*Test*
Baseline systems						
DEEPSPECTRUM	1024	0.4139 (0.3433 ± 0.0548)	**0.4239**	0.5741 (0.5395 ± 0.0207)	**0.4931**	**0.4585**
EGEMAPS	88	0.4112 (0.3168 ± 0.0459)	0.2975	0.5090 (0.4744 ± 0.0244)	0.3988	0.3482
BPM + ECG + RESP	3	0.3917 (0.2793 ± 0.0782)	0.1095	0.4361 (0.2906 ± 0.0787)	0.1861	0.1478
Proposed systems						
Physiological						
UCLBS	20	0.1606 (0.1356 ± 0.0149)	0.0794	0.3994 (0.3286 ± 0.0410)	**0.3044**	0.1920
RESP-CNN+ECG-CNN+BPM-CNN	350	0.4315 (0.3899 ± 0.0442)	0.1340	0.5445 (0.5323 ± 0.0130)	0.1814	0.1577
UCLBS+ECG-CNN+BPM-CNN	295	0.4333 (0.3749 ± 0.0421)	**0.1890**	0.5794 (0.5505 ± 0.0219)	0.2595	**0.2242**
Acoustic						
RAW(SER)	20	0.3404 (0.2986 ± 0.0311)	0.4338	0.5548 (0.5403 ± 0.0116)	**0.5134**	0.4736
RAW(PHN)	2048	0.3515 (0.3371 ± 0.0102)	**0.4909**	0.4122 (0.3894 ± 0.0217)	0.4767	0.4838
RAW(SER)+RAW(PHN)	2068	0.3742 (0.3540 ± 0.0176)	0.4850	0.4081 (0.3804 ± 0.0214)	0.4966	**0.4908**
Multi-modal early fusion						
UCLBS+Raw(SER)	40	0.4382 (0.3700 ± 0.0506)	0.3218	0.5602 (0.5273 ± 0.0222)	0.3597	0.3407
UCLBS+RAW(PHN)	2068	0.3803 (0.3579 ± 0.0189)	0.4644	0.4529 (0.4027 ± 0.0258)	**0.4952**	**0.4798**
RAW(SER)+RAW(PHN)+DEEPSPECTRUM	3092	0.3764 (0.3490 ± 0.0257)	**0.4734**	0.4280 (0.4114 ± 0.0195)	0.4386	0.4560

4.4.3.1 Uni-modal systems

Modelling acoustic signals: From Table 4.7, it could be observed that the proposed acoustic embeddings, RAW(SER) and RAW(PHN) are able to surpass the best performing DEEPSPECTRUM baseline results on the *Test* set for both valence and arousal. Also, the overall result obtained via these embeddings on the *Test* set are outperforming the best performing baseline systems. The hypothesis that the task-at-hand deals with speech emotion and SER based embeddings might help seems correct. Furthermore, it is interesting to observe that although our SER system was trained on IEMOCAP (Busso et al., 2008), an English corpora, the embeddings derived from it (RAW(SER)) generalised well for MuSe-Stress data which is recorded in German language. It is also worth noting that the RAW(SER) embeddings despite only being 20-dimensional, provide the best standalone results for emotional valence prediction.

Similar to RAW(SER), RAW(PHN) embeddings also appears to be robust towards unseen language, given that it generalises well for the MuSe-Stress data despite the network for deriving RAW(PHN) being trained on an English language corpora. These results also showcase that phonetic level information is crucial and complement emotion recognition task. It is worth mentioning that phonetic embeddings also showed good results for the case of non-speech vocalizations (Purohit et al., 2022) outperforming the DEEPSPECTRUM baseline for the ExVo multi-learning task (Baird et al., 2022). The RAW(PHN) embedding gives the best results for the emotional arousal prediction for a standalone embedding.

Moreover, it is interesting to see that the RAW(PHN) and RAW(SER) embeddings complement one another, with early fusion results of these embeddings providing a superior overall score.

Modelling physiological signals: From Table 4.7, we can see that all of our proposed physiological modelling methods outperform the physiological baseline from (Christ et al., 2022). The fusion of the features extracted from the proposed physiological CNNs (RESP-CNN+ECG-CNN+BPM-CNN) improves *Test* performance over the baseline, signifying the viability of directly modelling raw physiological signals.

It is worth noting that the feature embeddings extracted from the breathing pattern estimation model (UCLBS), while performing slightly worse for arousal estimation in comparison to the baseline (0.0794 vs 0.1095), significantly outperforms both the baseline and the proposed physiological CNNs for valence as well as the combined *Test* CCC performance. A possible explanation for this phenomena is the fact that these embeddings are extracted from a pre-trained network trained on raw waveforms from a conversational speech database, and thus potentially include speech related discriminative information which have been demonstrated to be informative for other tasks (Mostaani, Prasad, et al., 2022; Mostaani and Magimai.-Doss, 2022), further boosting their viability. We also note that, similar to the systems used for modeling acoustic features this network is also pre-trained on English database and is generalizing well on MuSe-Stress data which is in German language.

Given the better performance of UCLBS features, we decided to replace the RESP-CNN embeddings and training a fusion of UCLBS features with the other physiological CNNs (UCLBS+ECG-CNN+BPM-CNN), which, by trading off the excellent valence performance of UCLBS features for a significant improvement in arousal performance results in a 50% relative improvement in combined *Test* score over the physiological-only baseline (0.2242 vs 0.1478). However, it is worth noting that there is still a very large discrepancy between *Dev* and *Test* performance for physiological-only systems, highlighting that these systems struggle with overfitting on training distribution.

4.4.3.2 Multi-modal systems

Following results of uni-modal systems, we experiment with select multi-modal early fusion approaches of the top performing systems across modalities. First, we fused UCLBS breathing estimation features, which were the best performing standalone physiological feature set, with RAW(PHN) and RAW(SER) features, which are our top performing uni-modal features. However, the subsequent fused features showed performance degradation over the constituent acoustic features, with a much larger degradation observed for RAW(PHN). We also fuse RAW(SER) and RAW(PHN) with the best baseline feature representation (DEEPSPECTRUM), which also does not improve the performance.

4.4.4 Summary of the study

For MuSe 2022 stress sub-challenge, we investigated modeling of different feature embeddings obtained from task specific pre-trained neural networks, as well as modeling of physiological

signals for the continuous estimation of arousal and valence. Multi-modal systems were investigated by using early fusion between different modalities. Additionally, we investigated modeling of physiological signals in an end-to-end manner for the task-at-hand. While the proposed physiological models outperform the physiological baseline (Christ et al., 2022), embeddings extracted from pre-trained networks perform much better for valence and arousal estimation, including the pre-trained breathing pattern embeddings that model speech signals, demonstrating that acoustic features tend to be more informative for valence and arousal estimation when compared to physiological related information.

4.5 Conclusion

In this chapter, we built upon the study of breathing pattern estimation from raw waveform speech signals presented in Chapter 3. We investigated the utility of embeddings extracted from the pre-trained BPE CNNs for three applications.

First, we used the developed BPE networks for COVID-19 detection. COVID-19 is a respiratory disease that can potentially affect the breathing of a person. We compared the performance of BPE embeddings with embeddings extracted from a pre-trained PHR network as well as knowledge-based ComParE LLDs and observed that they provide better classification performance compared to the knowledge-based features but are outperformed by the embeddings extracted from the pre-trained PHR network. The second application involved distinguishing between natural and synthetic speech. As it is unclear whether synthetic speech carries similar breathing related information as natural speech, the inclusion of breathing pattern information could potentially be used in speech technologies such as presentation attack detection. We observed that the embeddings extracted from the BPE networks can distinguish between natural and synthetic speech signals if the synthetic speech signals are generated using TTS systems. In contrast, if VC methods are used for generating synthetic speech signals, the embeddings are not able to classify between natural and synthetic speech signals. For our last study, we investigated the usability of BPE networks for emotion recognition. Studies have shown that respiration could be altered based on emotional state of a person. We compared the performance of BPE embeddings with embeddings extracted from pre-trained PHR and SER networks as well as end-to-end models which model raw physiological signals for emotion recognition. We observed that the physiological information extracted by the BPE network provided complementary information to the speech related features for emotion recognition.

5 Analysis of the Breathing Pattern Estimation Networks

5.1 Introduction

In chapter 3 we developed convolutional neural networks for estimating breathing pattern from raw speech waveforms. We showed that the embeddings extracted from such networks can be used in different downstream tasks such as COVID-19 detection, distinguishing between natural and synthetic speech, and emotion recognition in chapter 4. In this chapter we investigate the networks more deeply to understand the type of information that is being modeled by them.

Both raw audio signals and spectral features were used to estimate breathing pattern from speech signals in Chapter 3. In Section 3.6.2 we demonstrated that the neural networks trained using log Mel spectrograms tend to model the spectral envelope which is covering the entire bandwidth while the CNNs trained using raw waveform signals focus on lower frequency regions. We decreased the frequency range modeled by the spectral based method to around 2.0 kHz by decreasing the number of Mel filter banks from 40 to 20 and observed a drop in the performance of the system.

Unlike the short-term speech processing based approaches, the end-to-end CNNs are opaque. In the sense that, while they achieve competitive breathing pattern estimation (BPE) performance, it is not clear what kind of information they capture or model for breathing pattern estimation. Figure 5.1 shows the cumulative frequency response of the first convolution layer of two raw waveform based CNNs, one trained on Philips database and one trained on UCL-SBM database for estimating breathing pattern from speech signals. The input window size for the models is 3 seconds, and they are trained using MSE loss function. As it can be seen, the trained CNNs primarily emphasize low frequency regions, though they also capture certain high frequency components. The cumulative frequency response provides insights into the filtering characteristics of the first convolutional layer. However, it remains unclear which specific frequency regions are predominantly modeled by the raw waveform based BPE networks as a whole.

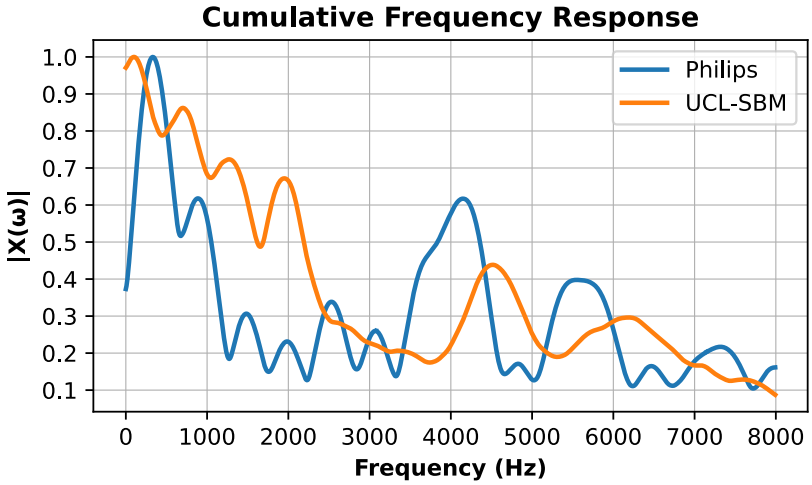


Figure 5.1 – The cumulative frequency response of the first layer in the CNNs pre-trained on Philips and UCL-SBM database when the input window is 3 seconds.

In this chapter we probe the raw waveform based BPE networks to gain insight into the type of information that is being modeled by them. We investigate the behaviour of the BPE networks when the frequency content of the speech signal is limited. Our aim is to understand the importance of different frequency regions and their relation to underlying physiological events namely source excitation and system articulation.

The rest of the chapter is organized as follows. We present the study design in Section 5.2 and the experimental setup is presented in section 5.3. The results are demonstrated in Section 5.4. Finally, we conclude in Section 5.6. The material presented in this chapter has not yet been published elsewhere.

5.2 Study Design

We chose several raw waveform based CNN models introduced in Chapter 3 which were trained on both Philips and UCL-SBM databases (see Sections 3.3.2 and 3.4). Our aim is to probe these pre-trained models to investigate how their prediction changes when the frequency content of the speech signal is limited. Our hypothesis is that if we remove a frequency region that is important for the BPE network, the output of the network will deviate from the original breathing pattern while if we remove a frequency region that is not important for the BPE network, the output of the networks will remain close to the original breathing pattern. Ideally we would like to have access to the ground truth breathing pattern for studying the effect of the frequency content of the speech signal on the BPE performance. In the presence of ground truth values, it is possible to compute relevance signals using gradient based methods as presented by (Muckenhirn et al., 2019). However, this has not been possible due to the lack of access to such data. Therefore, we design our study based on two downstream tasks that have been proven to be successful with the embeddings extracted from the BPE

networks in Chapter 4, namely COVID-19 detection and distinguishing between natural and synthetic speech.

We systemically decrease the bandwidth of the speech signal and analyse the deviation in the (a) output breathing pattern, (b) the neural embeddings and (c) the performance of the neural embeddings used in two aforementioned downstream tasks without any fine-tuning of the pre-trained models. If the BPE networks rely on a specific frequency region for their predictions, removing that frequency region should result in a deviation from the original breathing pattern. This deviation should be reflected in the embeddings extracted from the neural networks and the performance of the embeddings in the downstream tasks.

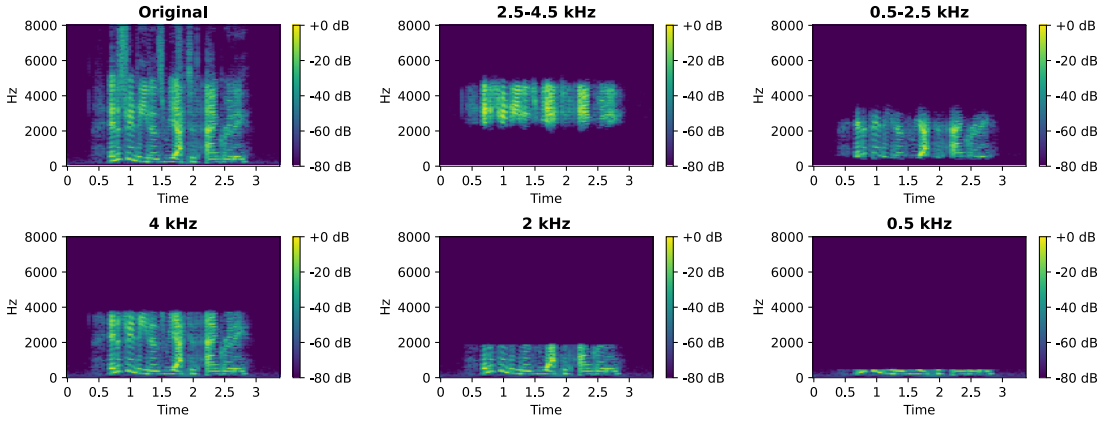


Figure 5.2 – The spectogram for an example audio from ASVspoof2019 database processed with two approaches. The top left is the original audio with maximum available frequency of 8 kHz. The top middle and right are examples of the filtered audio. The bottom line is examples of downsampled audio signals to obtain audio with maximum available frequencies of 4, 2, and 0.5 kHz.

In order to decrease the bandwidth of the speech signal we follow two approaches:

1. **Downsampling:** We downsample the speech signal to 8, 4, 2, and 1 kHz and then upsample the speech signal back to 16 kHz to be able to use them with the pre-trained neural networks which require input speech signals with sampling rate of 16 kHz. Like this we effectively limit the frequency content of the speech signals to 4, 2, 1, and 0.5 kHz respectively. In this approach we always keep the very low frequency content which could include source excitation related information. From here onward we denote the downsampled and upsampled audio signals with their highest available frequency content. For example an audio signal denoted as 4 kHz refers to an audio that has been downsampled to 8 kHz and then upsampled to 16 kHz. *Librosa* python library (McFee et al., 2015) is used for processing the audio signals.

2. **Bandpass filtering:** We apply bandpass filters with cut-off frequencies of 0.5-2.5 kHz, 1-3 kHz, 2.5-4.5 kHz, 4.5-6.5 kHz. The processed audio, while being limited in the frequency content has sampling rate of 16 kHz and can be used as input to the pre-trained neural

networks. With this approach, we can remove the very low frequencies in the speech signal and investigate their importance for the BPE networks. A butterworth fitler of degree 6 from *Scypi* python package (Virtanen et al., 2020) is used for processing the audio signals.

Figure 5.2 shows the spectogram of the original and processed audio for an example file. It can be seen that the proposed approaches effectively constrain the frequency content of the speech signal.

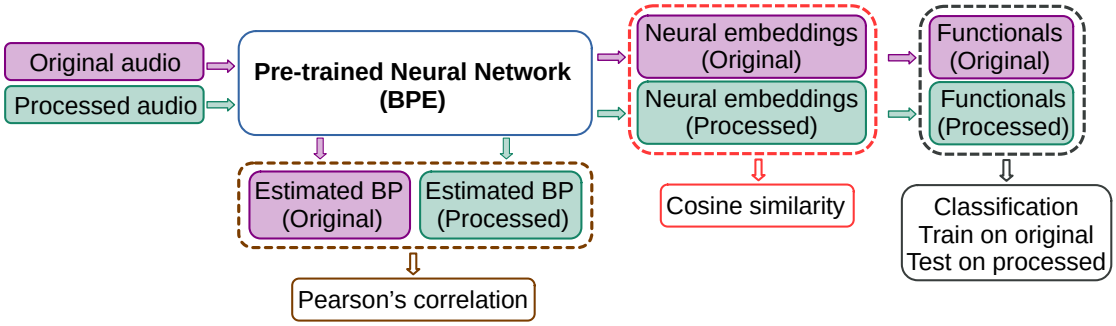


Figure 5.3 – An overview of the study design. The output and the embeddings extracted from the layer before the output of BPE networks are investigated, as well as the performance of the embeddings in downstream tasks. BP denotes breathing pattern.

Figure 5.3 presents the pipeline of our study. We investigat the behaviour of the BPE networks when the input speech signal has limited frequency content in three different ways:

- **Output:** We compare the estimated breathing pattern of the CNNs when the input signal is the original or the processed audio signals. We calculate the Pearson's correlation coefficient (r) between the two outputs as objective measure of similarity.

- **Embeddings:** We extract the embeddings from the neural networks for the original and processed audio signals. We calculate the cosine similarity between the embeddings as a measure of similarity.

- **Classification:** We use the embeddings extracted from the CNNs in two downstream tasks, COVID-19 detection and distinguishing between natural and synthetic speech which were previously used in Chapter 4 (see Sections 4.2 and 4.3). We follow the same pipeline as previously described by calculating the functionals (mean and standard deviation) for the embeddings extracted from the CNNs and training a Random Forest (RF) classifier on them (see Sections 4.2.1 and 4.3.1). We train the classifier on the embeddings extracted from the original audio signals and test them on the embeddings extracted from the processed audio signals and report the area under the receiver operating characteristic curve (AUC) as the evaluation measure.

5.3 Experimental Setup

In this section we first present the pre-trained BPE networks used in our study. We then present the databases of the chosen auxiliary tasks. Finally we present the similarity measures used and the details for the classification task.

5.3.1 Breathing pattern estimation networks

We chose 4 different raw waveform based CNN models trained on the Philips database which contains read speech and UCL-SBM database which contains conversational speech (see Sections 3.3.2 and 3.4) without any fine-tuning. They take 2 and 3 seconds of speech as input and are trained with MSE as a regression loss function. These models have been used in Section 4.2 and 4.3 for the COVID-19 detection and distinguishing between natural and synthetic speech respectively.

5.3.2 Databases

ASVspoof2019: This database has been provided by the automatic speech verification community to advance the presentation attack detection technologies. The goal for a security system is to distinguish between a genuine sample and an artificially created sample to imitate another person. We use the same portion of the database as in Section 4.3.2.1 in which the attacks are generated using Text-to-speech (TTS) and Voice conversion (VC) technologies. The database includes three subsets for training, development, and evaluation. We develop our methods by accumulating the training and development set and report the results on the evaluation set. This is the reporting set for the ASVspoof2019 dataset in this chapter, hereafter mentioned as *Test* set.

DiCOVA-II: This database, provided in the context of the second DiCOVA challenge, was provided to develop methods for detecting COVID-19 positive cases using speech sounds. The dataset includes training and development data that is organized in a 5-fold cross validation setting. Models are trained on the training set for each fold and the performance of the system is then reported on the accumulation of the development set from all the folds similar to Section 4.2.2.1. This is the reporting subset for the DiCOVA-II dataset in this chapter, hereafter mentioned as *Test* set.

5.3.3 Similarity measures

Output: We calculate the Pearson's correlation coefficient (r) of the estimated breathing pattern between the output of the system with maximum available frequency of 8 kHz as reference and all the other processed audio signals (downsampled and filtered) with limited frequency contents. The reported r value is the average Pearson's correlation coefficient for all the utterances in the *Test* set.

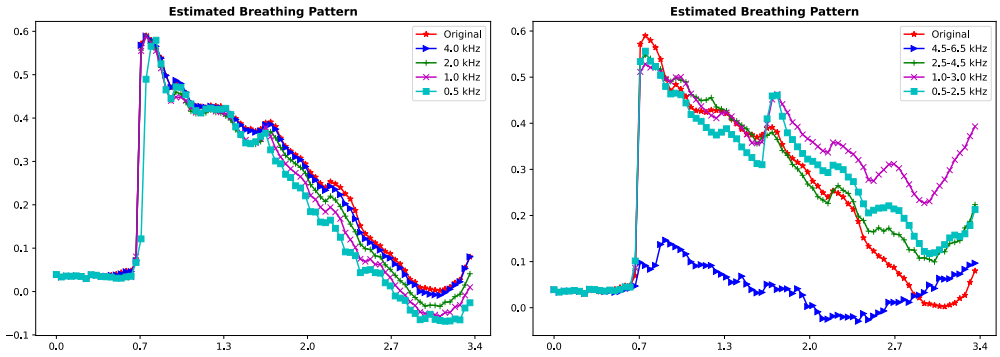
Embeddings: We extract the embeddings from the CNNs using the original and processed speech signals. For each utterance we have a 10 dimensional vector (f_01 - f_10). We look at each feature dimension separately over all the *Test* set and calculate the cosine similarity (CS) between the normalized histogram of the embeddings from the original speech signal as a reference to all the other processed audios.

5.3.4 Classification

We follow similar classification pipeline as presented in Section 4.3.2.3 for ASVspoof2019 database and in Section 4.2.2.3 for DiCOVA-II database. Fixed-length representation of the embeddings for each utterance is obtained using functionals (mean and standard deviation). The resulting 20-dimensional representations are then used with a Random Forest (RF) classifier (Ho, 1995) and the evaluation measure is reported over the *Test* set. Unlike the work in Chapter 4 we do not tune any hyperparameters and use the default values of the *Scikit-learn* (Pedregosa et al., 2011) toolkit for classification. The same classification framework is deployed for both the original and the processed audio signals.

5.4 Results

5.4.1 Output



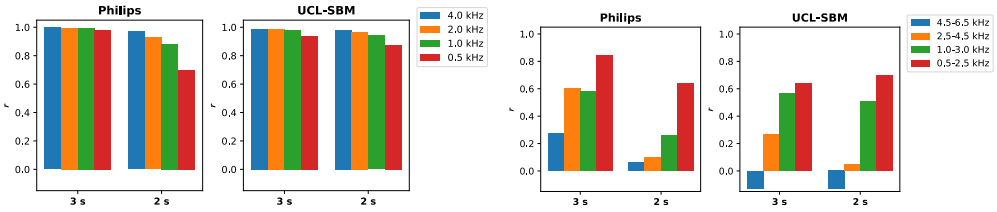
(a) – Downsampled **(b)** – Filtered

Figure 5.4 – The estimated breathing pattern from a model pre-trained on Philips database with 3 seconds of input data for (a) downsampled and (b) filtered signal. The input data is from the ASVspoof2019 database.

Figure 5.4 shows the estimated breathing pattern from an example audio file with different downsampling (Figure 5.4a) and filtering (Figure 5.4b) methods. It can be seen that as the signal is progressively downsampled, and the highest available frequency decreases while the very low frequency content of the speech is kept, the output from the BPE networks systematically deviates from the original signal. We observe a deviation in the output of BPE networks for filtered signals compared to the original signal as well, but, in this case, no clear pattern of deviation is observed in relation to the cut-off frequencies of the bandpass filters

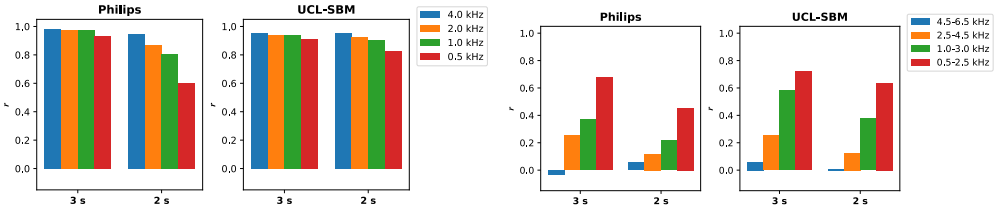
with one exception. When the signal is filtered to retain only high frequency components (4.5–6.5 kHz), the deviation is much larger.

This observation is further supported by the Pearson's correlation coefficient (r) between the output of the original and processed audio signals. Figures 5.5 and 5.6 illustrate the r value for the *Test* set of ASVspoof2019 and DiCOVA-II databases, respectively. A consistent pattern emerges across both databases, and for models trained on both Philips and UCL-SBM databases. As the downsampling increases, the r value decreases, with the effect being particularly pronounced for a 2 second input window. When the signal is filtered, the drop in r value is higher, with the most substantial reduction occurring when the signal is filtered to retain only high frequency content.



(a) – Downsampled **(b)** – Filtered

Figure 5.5 – The correlation between the output of the original and (a) downsampled and (b) filtered signal. The input data is from the ASVspoof2019 database.



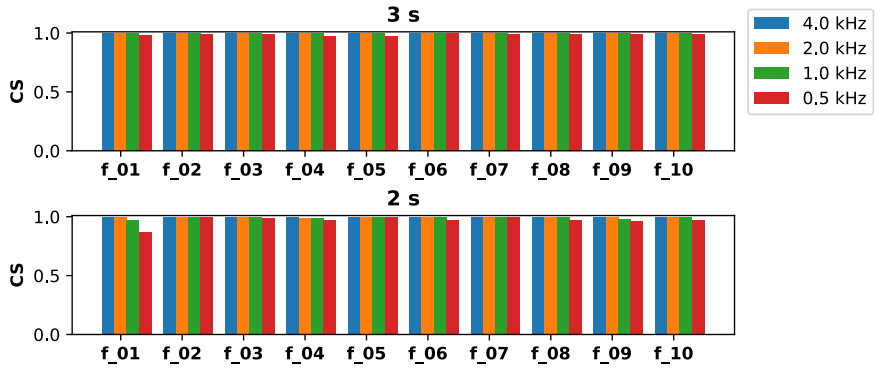
(a) – Downsampled **(b)** – Filtered

Figure 5.6 – The correlation between the output of the original and (a) downsampled and (b) filtered signal. The input data is from the DiCOVA-II database.

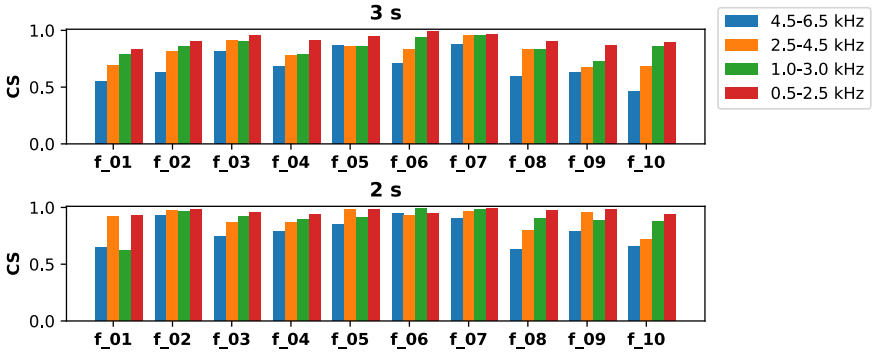
5.4.2 Embeddings

The embeddings extracted from neural networks are used in various downstream tasks. Therefore, it is worth investigating the effect of limiting the frequency content of the speech signal on these embeddings. Figure 5.7 shows the cosine similarity (CS) between the embeddings extracted from the neural networks for the original and processed audio signals. The input audio signals are taken from the ASVspoof2019 database, and the models are pre-trained on Philips database. The input windows are 2 and 3 seconds in duration. It can be seen that downsampling the signal decreases the similarity between the embeddings. This is more prominent when the input window is 2 seconds. Similar to the observations in the output, the results indicate that downsampling the signal decreases the similarity between the embeddings, with this effect being more pronounced for the 2 second input window. The deviation

in the embeddings becomes more substantial when the signal is filtered, particularly when only high frequency components are retained.



(a) – Downsampled



(b) – Filtered

Figure 5.7 – The cosine similarity between the output of the original and (a) downsampled and (b) filtered signal. The input data is from ASVspoof2019 database. The models are pre-trained on Philips database.

5.4.3 Classification

We examine how the deviation from the original embeddings impacts the performance of classifiers in downstream tasks. Table 5.1 shows the AUC on the *Test* set of ASVspoof2019 database for an RF classifier to distinguish between natural and synthetic speech signals. The performance remains stable when the input signal is downsampled, as long as the correlation between the original and processed signals remains high. When the signal is filtered, the performance deviates more substantially from the original system's performance. This is expected because as the output deviates more from the original signal, it is unclear what specific information the networks are modeling. Additionally, differences in frequency content between training and testing data introduce further uncertainty, especially when the correlation between the breathing patterns from original and processed signals are low. A similar pattern is observed in the COVID-19 detection task using the DiCOVA-II database, as shown in Table 5.2.

Table 5.1 – The performance of RF classifier on ASVspoof2019 database using BPE networks pre-trained on Philips and UCL-SBM databases.

	AUC								
	Processed audio with limited frequency content (kHz)								
	original	Downsampled				Filtered			
	8	4	2	1	0.5	4.5-6.5	2.5-4.5	1.0-3.0	0.5-2.5
Models trained with 3 seconds of input									
Philips	88.75	88.77	88.76	88.31	85.75	57.03	59.92	76.06	77.19
UCL-SBM	89.86	89.51	89.57	89.43	86.29	64.06	74.38	65.07	81.8
Models trained with 2 seconds of input									
Philips	88.1	88.26	88.16	87.87	86.09	62.53	73.11	69.03	84.69
UCL-SBM	88.39	88.1	87.64	86.78	83.41	57.18	78.63	72.13	84.75

Table 5.2 – The performance of RF classifier on DiCOVA-II database using BPE networks pre-trained on Philips and UCL-SBM databases.

	AUC								
	Processed audio with limited frequency content (kHz)								
	original	Downsampled				Filtered			
	8	4	2	1	0.5	4.5-6.5	2.5-4.5	1.0-3.0	0.5-2.5
Models trained with 3 seconds of input									
Philips	67.04	66.09	65.41	65.84	65.82	54.1	53.95	59.41	57.55
UCL-SBM	64.75	64.08	64.43	62.4	59.73	54.84	54.3	47.6	62.84
Models trained with 2 seconds of input									
Philips	65.75	66.61	64.63	62.93	60.75	53.97	50.62	55.45	56.12
UCL-SBM	60.81	61.2	60.18	59.22	63.45	53.41	49.47	57.8	52.91

5.5 Discussion

The results presented in Sections 5.4.1, 5.4.2, and 5.4.3 suggest that the raw waveform based BPE networks rely on specific frequency regions for their predictions. Removing these regions leads to deviations from the original breathing patterns, which are reflected in the embeddings extracted from the neural networks and in the performance of these embeddings in downstream tasks as well.

Our experiments reveal that BPE networks predominantly rely on very low frequencies (below 1 kHz) for their predictions when the input window is set to 3 seconds. When the signal is downsampled to 1 kHz, the Pearson's correlation coefficient (r) remains close to 1. However, downsampling below 1 kHz results in a noticeable drop in r. For example, for a model pre-trained on the Philips database with a 3 second input window, the correlations shown in Figure 5.5a are 0.9977 and 0.9941 for downsampling to 4 kHz and 1 kHz, respectively while downsampling to 0.5 kHz results in a relatively lower correlation of 0.9798. Additionally, r values are higher when the signal is downsampled to 2 kHz (retaining frequencies below

0.5 kHz) compared to when the signal is filtered to retain only frequencies between 0.5–2.5 kHz (excluding frequencies below 0.5 kHz). For instance, for the same model pre-trained on the Philips database, the correlation drops significantly from 0.9962 to 0.8442 when the frequencies below 0.5 kHz are removed. This suggests that frequencies below 0.5 kHz are particularly critical for the performance of BPE networks. These observations indicate that the networks are modeling information related to vocal fold excitation (the source). Vocal fold excitation, which is associated with pitch, exhibits the highest energy in frequencies below 0.5 kHz. This likely explains the critical importance of frequencies below 0.5 kHz in the speech signal for these networks.

The duration of the input signal also appears to influence the networks' performance. The r value is lower for a 2 second input window compared to a 3 second input window, with a more pronounced drop as higher frequencies are removed. For instance, in Figure 5.5a, for a model pre-trained on the Philips database, the r value decreases from 0.9977 for a 3 second input window to 0.97 for a 2 second window when the signal is downsampled to 4 kHz. When the signal is downsampled to 0.5 kHz, the r value drops further, from 0.98 to 0.70. This suggests that when the input duration is shorter, the BPE networks begin to rely on the higher frequency content of the speech signal as well. Consequently, removing higher frequencies has a greater impact on the networks' output with shorter input durations.

These observations may be related to the phonetic content of the speech signal. Most vowels have their first and second formants below 2.5 kHz, where their energy is concentrated. Additionally, voiced consonants which require the vibration of the vocal folds also have energy in low frequency regions. Moreover, vowels are usually longer than consonants. In other words, they represent a larger portion of acoustic events in speech signals. This could explain why the networks tend to rely on these lower frequencies for their predictions. When the input duration is very short, the networks may lack sufficient information for accurate predictions and thus depend more on higher frequency regions with lower energy content. It is worth mentioning that a recent work by (Nallanthighal, Harma, et al., 2021) demonstrated that respiratory effort, including lung volume changes, is influenced by the phonetic content of speech. Our analysis suggests that raw waveform based CNN models are modeling phonetic level acoustic information for breathing pattern estimation.

5.6 Conclusion

In this chapter, we explored the types of information captured by raw waveform based BPE networks by systematically limiting the frequency content of speech signals and analysing the resulting network outputs, neural embeddings, and classification tasks using these embeddings. We applied two methods — downsampling and bandpass filtering — to isolate specific frequency regions within the speech signals. The first approach retained frequencies below 0.5 kHz, while the second excluded them. We examined models trained on the Philips and UCL-SBM databases, using input windows of 2 and 3 seconds, across two downstream tasks:

COVID-19 detection and distinguishing between natural and synthetic speech.

The processed signals were passed through the BPE networks, and their outputs, embeddings, and classification accuracy were compared to those obtained from the original signals. Our findings indicate that the BPE networks rely primarily on low frequency content of speech signal, particularly frequencies below 0.5 kHz, for accurate predictions. However, when the input duration was shortened from 3 seconds to 2 seconds, the networks started also relying on higher frequency regions. Notably, high frequencies alone (above 4.5 kHz) proved insufficient for accurate predictions. This observation can be linked to phonetic level acoustic information, as both vowels and voiced consonants contain energy in low frequency regions, and phonation involves airflow through the vocal folds. Additionally, the first formants of vowels, which generally have high energy, appear within low frequency regions. This suggests that BPE networks are modeling both source-related information, such as pitch, and phonetic-level details for predictive accuracy.

6 Cardiac Activity and Speech

6.1 Introduction

In previous chapters we developed models to extract breathing related information from speech signals, applied these pre-trained models to various applications, and investigated the nature of the information being modeled. In this chapter, we shift our focus to studying the relationship between cardiac activity and speech signals.

There has been an effort to examine the association between speech signals and cardiac functions. (Orlikoff et al., 1989) in one of the very early studies showed that cardiovascular system can influence the vocal fundamental frequency (F0) indicating that the absolute F0 perturbation (jitter) during a sustained phonation could vary between 0.5% to 20%. In other studies, cardiac activity was studied in relation with speech in different emotional states. (Williams et al., 1972) demonstrated that the emotion variation might cause an increase in blood pressure (BP), heart rate (HR), sub-glottal pressure, and the depth of respiratory movements. (A. P. James, 2015) suggested a strong correlation between speech, emotion, and heart rate using spectral features from speech. (J. Smith et al., 2017) showed that HR increases when the person is speaking compared to when they are silent and this increase is greater when they are frustrated. In another study (Ryskaliyev et al., 2016), the HR was predicted using linear models in different emotional states. (Jati et al., 2018) predicted physiological signals from speech during stressful conversations.

In these studies, the influence of inter- and intra-individual variability on the efficacy of speech based models remains relatively unexplored. For example in (Schuller et al., 2014; Mesleh et al., 2012; J. Smith et al., 2017; Usman et al., 2021), it has been shown that acoustic features, in particular spectral features, can be used as good predictors of cardiac activity parameters like beats per minute (BPM) in the context of regression and classification tasks. However, the use of speaker dependent data splits may have confounded the reported performances, as data from the same speakers was included in both training and testing sets. (Schuller, Friedmann, et al., 2013) addressed speaker dependency in their experimental design, demonstrating that a speaker dependent split outperformed a leave-one-speaker-out (LOSO) approach, though

the extent of inter- and intra-speaker variability was not thoroughly examined.

In this thesis, we aim to predict cardiac activity parameters such as BPM from speech signals while considering the impact of inter- and intra-individual variability on model performance. One challenge is the limited availability of large speech corpora that include physiological recordings, leading to models that overfit to specific speaker data. The scarcity of such datasets also complicates the assessment of the models' generalizability. To address these challenges, we employ the following approaches:

- One way to overcome the limitation of data availability is to pre-train models on large scale corpora and evaluate the pre-trained models on downstream paralinguistic tasks with limited data (see Section 2.1.1.2). A major advancement in this field is the introduction of self-supervised models (SSMs) which are trained to optimize specific task objectives without the use of labeled data. They are trained on large datasets and have shown superior performance on paralinguistic tasks (Scheidwasser-Clow et al., 2022; Shor et al., 2020) as well as speech (Yang et al., 2021) and audio tasks in general (Turian et al., 2022). Our study is based upon these findings by evaluating the ability of a self-supervised model to predict cardiac activity parameters such as BPM and compare its performance to that of acoustic features.
- We investigate the impact of both inter- and intra-individual variability on model performance and generalizability. We also study the optimal context window length for speech in predicting heart rate and identify the salient features for this task.
- We introduce a novel database containing simultaneously recorded speech and physiological signals, utilizing this resource to validate our findings.

The remainder of the chapter is organized as follows. The study design is demonstrated in Section 6.2. The experimental setup is detailed in Section 6.3. Results and analysis is given in Section 6.3.4, and finally we conclude in Section 6.6. The material presented in this chapter related to Ulm-TSST database was originally published in modified form in (Elbanna et al., 2024). All the rest, including the introduction of a novel database denoted as TIPS has not yet been published elsewhere.

6.2 Study Design

Predicting heart rate (HR) from speech signal can be viewed as a regression problem wherein each context window of speech is mapped to a BPM value in the output. In this study, BPM values initially derived from recorded electrocardiogram (ECG) signals. A ground truth BPM is assigned to each context window by calculating the average BPM over the corresponding context window of the BPM signal. The speech features are extracted from the same context window and are used as input to a regression model (also see Section 2.1). The model perfor-

mance is reported using coefficient of determination (R^2) and Pearson's correlation coefficient. Figure 6.1 illustrates the training pipeline for predicting BPM from speech signals.

As presented in Section 2.1.1, two commonly used approaches to extract speech features are knowledge-based and data-driven methods. In this study, we investigate the performance of these two approaches in predicting BPM from speech signals. Specifically, we use the following feature sets:

- **Knowledge-based:** We extract two widely used feature sets, namely eGeMAPS and ComParE features, which have been shown to perform well in many paralinguistic tasks (see Section 2.1.1.1). ComParE low level descriptors (LLDs) were previously employed in Section 4.2.2.2 for COVID-19 detection.
- **Data-driven:** As previously noted, self-supervised models pre-trained on large datasets have proven beneficial for a range of downstream tasks, including the detection of individuals' emotional states (see Sections 6.1 and 2.1.1.2). Given the close relationship between stress and cardiac activity, we selected the Hybrid BYOL-S model, a self-supervised learning framework previously utilized in various audio tasks, emotion recognition, and analysing speech under cognitive and physical load (Elbanna, Scheidwasser-Clow, et al., 2022; Elbanna, Biryukov, et al., 2022). Details of this model are presented in Section 6.3.3.

Additionally, we explore the effect of context window size on the performance of our system using two databases.

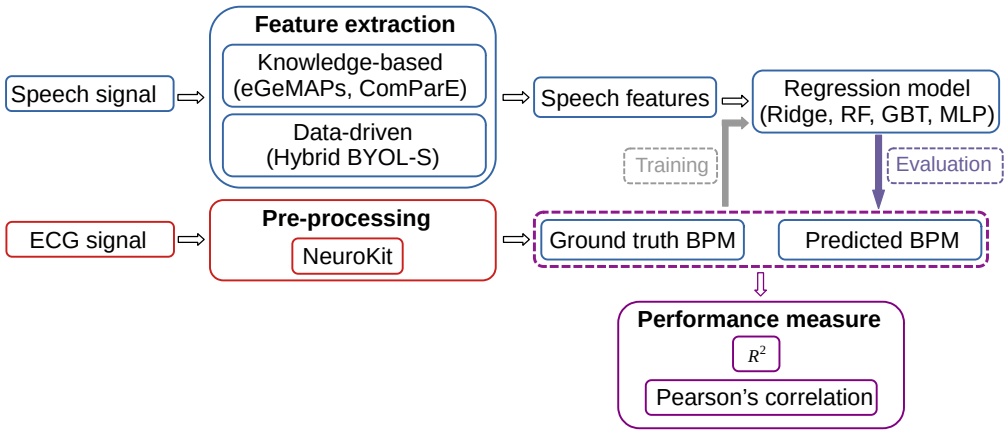


Figure 6.1 – Training pipeline for predicting BPM values from knowledge-based and data-driven speech representations.

6.3 Experimental Setup

In this section we introduce the databases and protocols employed in this chapter. We detail the procedure for extracting ground truth BPM and speech features, as well as the regression models utilized.

6.3.1 Databases

6.3.1.1 TIPS

The novel database, denoted as TIPS, was collected between June and July 2022, at the Idiap Research Institute, as part of this thesis work. It includes simultaneous recordings of speech and physiological data from healthy participants, enabling us to analyse the relationship between speech and physiological signals.

Physiological signals were recorded using a device named ICARUS, designed and manufactured by the Swiss Center for Electronics and Microtechnology (CSEM). The device consists of two sensors mounted on a belt worn around the chest (see Figure 6.2). It collects skin impedance (EDA), electrocardiogram (ECG), phonocardiogram (PHG), speech, and the movement data using an accelerometer with all signals recorded synchronously at variable sampling rates. In this study, we only use ECG signals recorded at a sampling rate of 250 Hz. simultaneously, speech signal with sampling rate of 44.1 kHz is recorded using a phone placed at approximately 20 cm from the participant's mouth. The aligned speech signals and ECG data collected by the ICARUS device are employed in subsequent analyses.

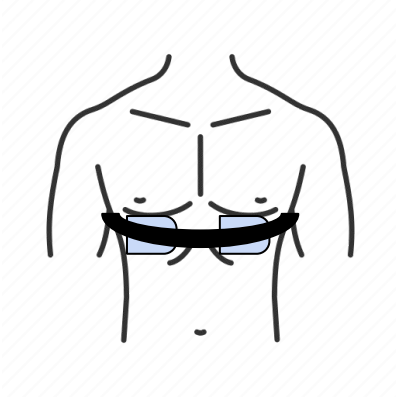


Figure 6.2 – The ICARUS device consists of two units mounted on a belt that is worn around the chest to record physiological signals.

During data collection, the participants performed four different tasks: (a) read speech, (b) free speech, (c) emotional sentences, and (d) picture reaction. This study focuses exclusively on the data obtained from the first two tasks. In the read speech task, participants were instructed to read aloud two texts - *The North Wind and the Sun* and *The Boy Who Cried Wolf*. In the free speech task, participants had the opportunity to discuss a topic of their choosing for approximately three minutes. Participants could choose to conduct the data collection in either English or French.

Data were collected from 32 participants (25 males and 7 females). 15 participants completed the data collection in English, while 17 did so in French. We segmented the speech and ECG signals according to the tasks and removed data from participants whose ECG signals were of

insufficient quality. The resulting database consists of 28 participants (22 males and 6 females) aged between 20 and 49 years. The participants are evenly split by language, with 14 speaking English and 14 speaking French. This led to a total of 56 recordings from the read speech task (approximately 104 minutes) and 28 recordings from the free speech task (approximately 67 minutes). The data from the read speech and free speech tasks are analysed separately, designated as TIPS-read speech and TIPS-free speech, respectively, throughout this chapter.

6.3.1.2 Ulm-TSST

The Ulm-TSST database, previously used in Section 4.4 for emotion recognition was part of MuSe-Stress sub-challenge of the MuSe challenges (Stappen, Baird, et al., 2021; Christ et al., 2022). It consists of data from 69 German speaking subjects who performed free speech tasks following the Trier Social Stress Test (TSST) protocol (Kirschbaum et al., 1993). Multiple physiological signals such as ECG, BPM, respiration (RESP), and EDA were captured during the speech tasks at a sampling rate of 1 kHz (also see Section 4.4.2.1). In this chapter we use the ECG signal to generate ground truth BPM and predict BPM values from speech signals.

6.3.2 Data preprocessing

Utterances in the Ulm-TSST database were originally acquired with 6 channels. We select the channel with the highest loudness (the first channel). The TIPS database was collected with only one channel. All mono-channel utterances are resampled to 16 kHz and standardized. Following this, we segment the data into clips of varying window sizes ranging from 3 to 5 seconds with a hop size of 500 ms. The ECG signals are preprocessed using *NeuroKit* package (Makowski et al., 2021). BPM is computed from the ECG signal using the same package. Lastly, we compute the average BPM value for each clip to have one value per audio sample.

6.3.3 Speech feature extraction

We compare the performance of knowledge-based features against speech representations generated from a pre-trained self-supervised model.

Knowledge-based: We extract two commonly used feature, namely eGeMAPS (88 features) (Eyben et al., 2016) and ComParE (6373 features) (Schuller, Steidl, Batliner, et al., 2013), using openSMILE (Eyben et al., 2010), an open source toolkit for extracting low-level descriptors from audio utterances (see Section 2.1.1.1).

Data-driven: We use Hybrid BYOL-S model (Elbanna, Biryukov, et al., 2022), a self-supervised model derived from Bootstrap Your Own Latent (BYOL-A) learning framework (Niizumi et al., 2021). BYOL-A learns general-purpose audio representations from juxtaposing two augmented views of a single input utterance. The two augmented views are fed to two networks, an

online and a target network. The task objective involves the online network predicting the generated representations from the target network. The new variant of BYOL-A (i.e. Hybrid BYOL-S) is a speech specific derivation that trains the online network to predict data-driven representations from the target network as well as knowledge-based representations (ComParE features) simultaneously yielding robust speech embeddings of size 2048. In this study, we use the Hybrid BYOL-S with the encoder replaced by a lightweight version of the convolutional vision transformer (CvT) (H. Wu et al., 2021) as in (Elbanna, Scheidwasser-Clow, et al., 2022; Elbanna, Biryukov, et al., 2022).

6.3.4 Experimental protocols

We evaluate the performance of the three feature sets, eGeMAPS, ComParE, and Hybrid BYOL-S, in predicting BPM. To assess the generalizability of the methods across various speakers, we use three data splits:

- **Speaker independent:** In this protocol, 70% of the speakers are used for training and the remaining 30% are used for testing. This ensures that the data from the same speaker is not present in both training and testing sets.

- **Speaker dependent:** Here, 70% of the clips for each speaker are allocated for training while the remaining 30% are used for testing, ensuring that data from all speakers is present in both training and testing sets.

- **Speaker specific:** In this case, regression models are trained on data samples from an individual speaker, with the training and testing samples split into 70% and 30%, respectively.

After training the regression models, the performance on the *Test* set is reported using coefficient of determination (R^2) and Pearson's correlation coefficient.

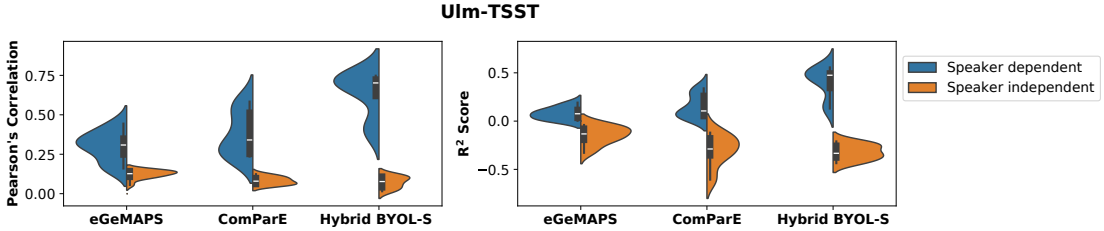


Figure 6.3 – Performance of different speech features in speaker dependent and speaker independent protocols for the Ulm-TSST database. The reported distributions show the evaluation across multiple regressors and window sizes.

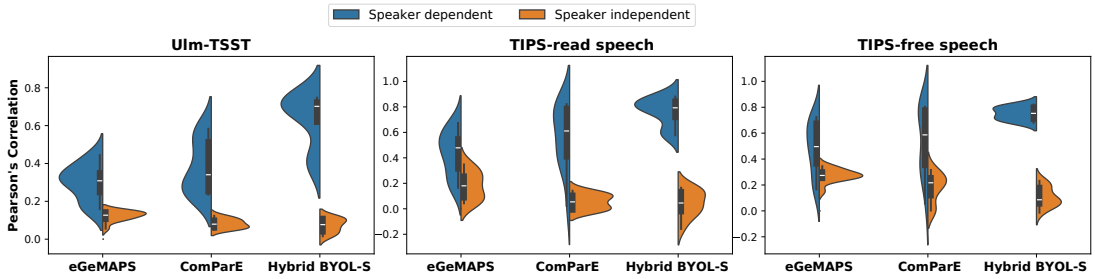


Figure 6.4 – Performance of different speech features in speaker dependent and speaker independent protocols across multiple regressors and window sizes for the three databases.

6.3.5 Regression models

For each experiment, multiple regressors are used such as Ridge regression, Random Forrest (RF), Gradient Boosting tree (GBT), and multi-layer perceptron (MLP) (see Section 2.1.2.1). A grid search is performed for hyperparameter optimization which is specific to each model. In case of speaker independent, a 5-fold group shuffle split is utilized where speakers in the training set are further divided into train and validation sets (70% and 30%, respectively). Thus, ensuring that the validation set includes unseen speakers during training. For speaker dependent and speaker specific experiments, 5-fold time series split is employed where the training samples for all speakers are split into train and validation (70% and 30%, respectively) while considering the temporal dependency between the data samples.

6.4 Results

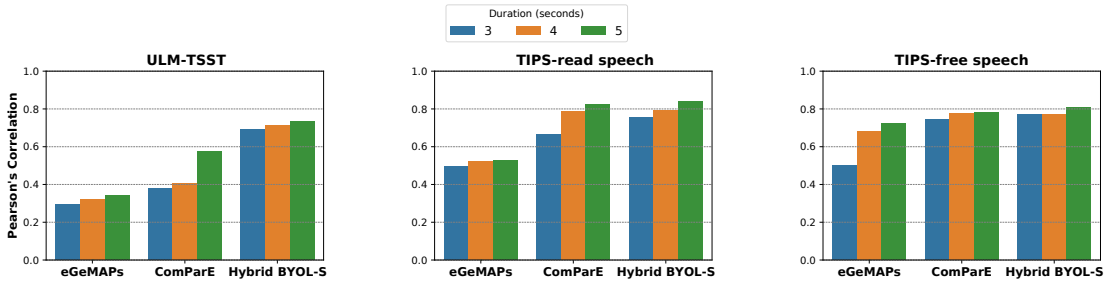


Figure 6.5 – Performance of different speech features with varying context window duration using three different datasets. The reported distributions show the evaluation using speaker dependent and GBT regression model for predicting BPM.

Figure 6.3 illustrates the performance of candidate speech representations on predicting BPM under two different protocols, speaker dependent and speaker independent, for Ulm-TSST database. We report the R^2 and Pearson's correlation between the predicted and the ground truth. The plotted distributions highlight the performance across different regressors and window sizes (i.e., 3, 4, and 5 seconds). It can be observed that Hybrid BYOL-S outperforms knowledge-based features in a speaker dependent setting. Whereas, all representations per-

form equally poorly in the speaker independent setting. The same trend is observed for the TIPS-read speech and TIPS-free speech databases as shown in Figure 6.4. This result highlights the limitations of generalizability of speech features for this downstream task.

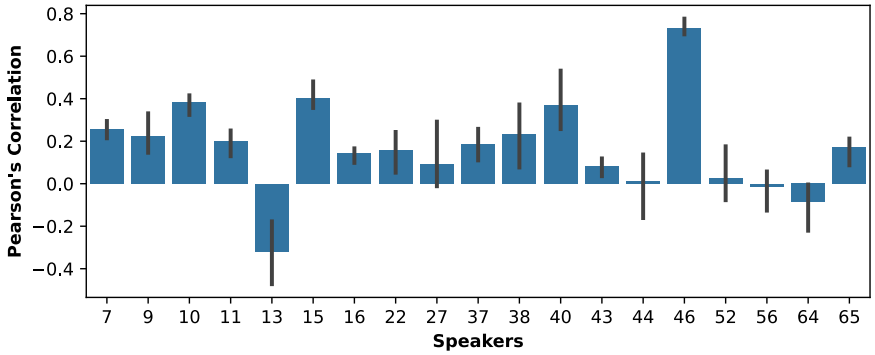


Figure 6.6 – Performance of Hybrid BYOL-S features for speaker specific protocol using 5 seconds of audio and a GBT regressor. The figure shows the obtained Pearson's correlation coefficient from a random set of 20 speakers chosen from Ulm-TSST dataset.

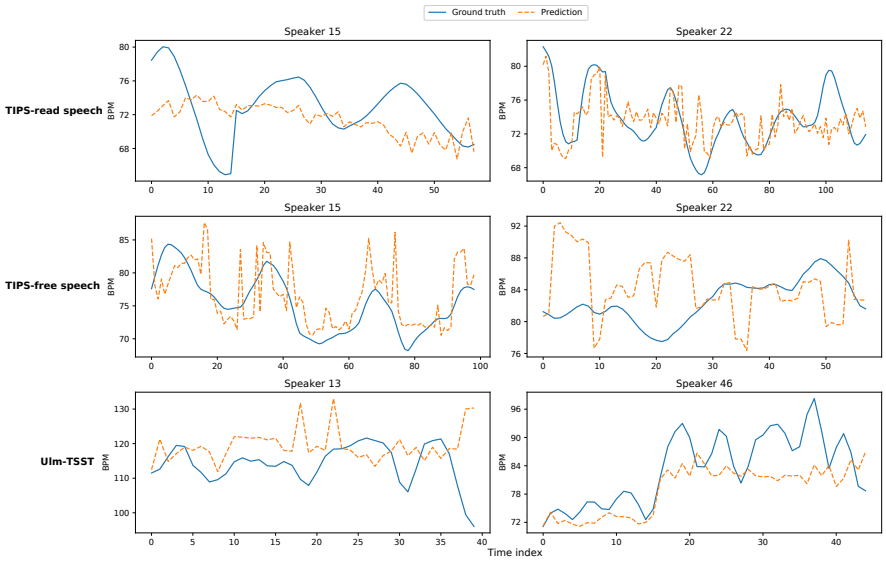


Figure 6.7 – Predictions from GBT model using Hybrid BYOL-S features with 5 seconds window size.

Moreover, we study the effect of context window duration on performance. Figure 6.5 reports the Pearson's correlation on a speaker dependent test set using GBT (best-performing regressor) for the three databasets. We observe that increasing the window size improves the performance in all feature candidates. Moreover, we observe considerable improvement between 3 second window and 4 second window while between 4 seconds and 5 seconds duration performance is relatively overlapping.

Given the discrepancy between both speaker related protocols, we further train regression

models on data samples from a single speaker resulting in building a model per speaker (i.e., speaker specific protocol). Figure 6.6 demonstrates the Pearson's correlation coefficient across a randomly selected sample of 20 speakers from Ulm-TSST database. The Hybrid BYOL-S is used to predict BPM across different window sizes. It can be seen that the performance varies considerably across speakers.

We further plot the predictions from two speakers from each dataset; as shown in Figure 6.7. It can be observed that while the model is able to capture the trend in BPM for some speakers, it fails to do so for others.

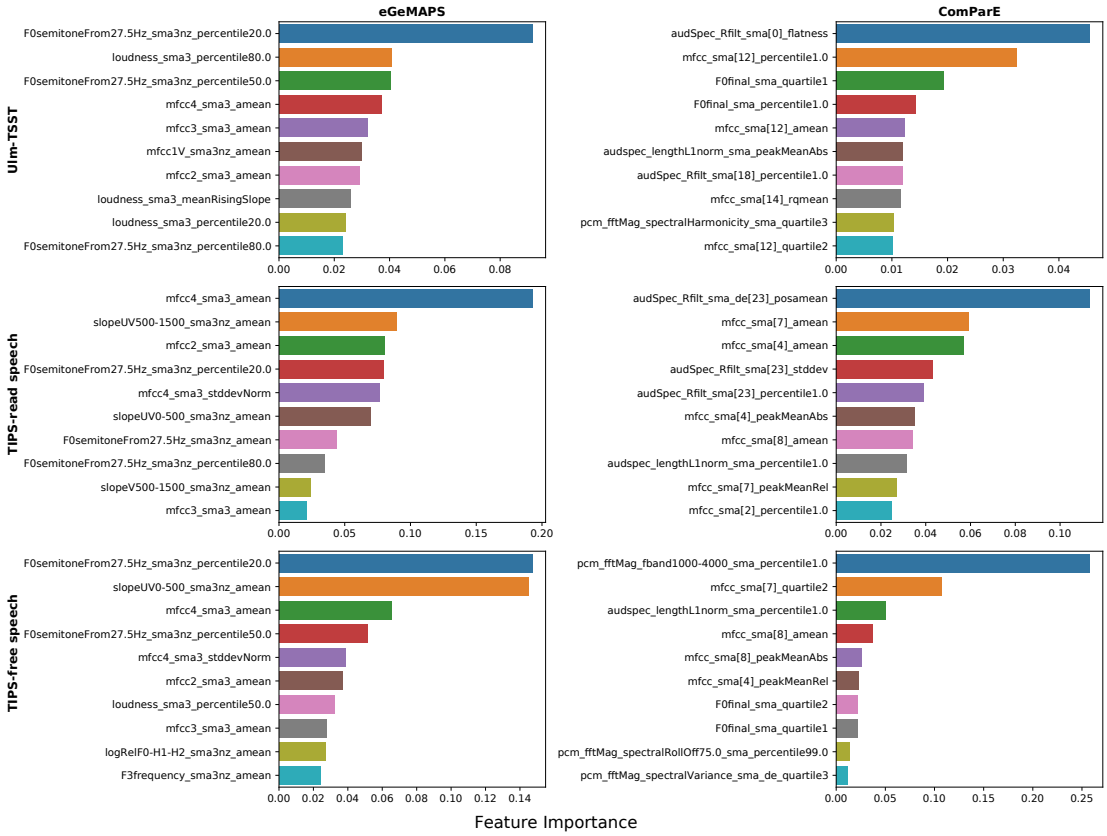


Figure 6.8 – Feature importance showing top 10 acoustic features for BPM extracted from both openSMILE feature sets, eGeMAPS and ComParE for the three data subsets. The feature names are denoted as *LLD_filtering_functional*

6.5 Analysis of Knowledge-based Features

Analyzing the eGeMAPS and ComParE feature sets provides valuable insights into the acoustic features that are important for predicting heart activity. To this end, we perform feature ranking on the eGeMAPS and ComParE features to identify the top 10 features contributing to the prediction of BPM. Feature importance is derived from the best-performing estimator (GBT) and ranked accordingly. Figure 6.8 illustrates the top 10 features from each feature set

that contribute to BPM prediction across the three datasets. A consistent trend is observed across all datasets. For eGeMAPS, features related to fundamental frequency (F0), loudness, and spectral characteristics emerge as key predictors. Similarly, for ComParE, spectral features and auditory spectral features are important for BPM prediction. These findings align with prior studies in the literature (Orlikoff et al., 1989; Mesleh et al., 2012; Schuller et al., 2014), which have explored the relationship between cardiac activity and acoustic features such as fundamental frequency and spectral features.

6.6 Conclusion

In this chapter, we explored the relationship between speech signals and cardiac parameters, with a focus on predicting heart rate from speech data. We introduced a novel database, TIPS, containing simultaneous recordings of both speech and physiological signals, including ECG. Participants in this study performed two types of speech tasks: read speech and free speech. In addition to TIPS, we utilized another database containing simultaneous speech and physiological signals previously employed in emotion recognition studies.

We investigated both knowledge-based and data-driven feature sets for predicting cardiac parameters. Our results demonstrated that Hybrid BYOL-S, a self-supervised model previously applied to stress detection, outperformed traditional knowledge-based features in predicting beats per minute (BPM) from speech signals. This underscores the potential of data-driven approaches, particularly self-supervised models (SSMs), in enhancing predictive accuracy.

However, we observed that all feature sets, including both data-driven and knowledge-based approaches, exhibited poor performance under a speaker independent protocol. This finding highlights the limitations of these features and their sensitivity to inter-individual variability. Similarly, we noted varying correlation scores when models were trained on a single speaker's data. Poor prediction accuracy for some speakers suggests that intra-individual variability also limits model generalization to unseen samples from the same individual.

Additionally, we demonstrated that increasing the context window duration to more than 3 seconds considerably improved prediction performance. A feature importance analysis revealed that spectral features and loudness were consistently the most critical acoustic features for this task. This analysis was validated across three different datasets, where we observed similar trends, indicating that our methodology is scalable and applicable to other datasets. Collectively, these results provide a deeper understanding of the robustness and limitations of both knowledge-based and data-driven features in the study of cardiac activity.

7 Hypoglycemia and Speech

7.1 Introduction

In previous chapters, we developed models to estimate breathing patterns from speech signals and explored their application in speech related tasks. We also investigated the relationship between cardiac function and speech. This chapter integrates these findings in a clinical context to examine how hypoglycemia affects the speech signals of diabetic patients.

Diabetes mellitus (DM) is a complex metabolic disorder characterized by elevated blood glucose levels due to insufficient insulin production (type 1 diabetes mellitus (T1DM)) or improper insulin use (type 2 diabetes mellitus (T2DM)). 537 million people were reported to have diabetes worldwide in 2021 and this number was projected to increase to 643 million by 2030, underscoring its global prevalence (Magliano et al., 2021). Effective management of diabetes requires diligent monitoring of blood glucose levels to prevent complications like hypoglycemia (low blood glucose) and hyperglycemia (high blood glucose). However, many diabetics often neglect regular monitoring due to the discomfort and cost of frequent blood tests, leading to serious health issues (Sidorova et al., 2022). This highlights the need for non-invasive and reliable alternatives.

The speech production system involves intricate interactions among the nervous, muscular, respiratory, and cardiovascular systems. Diabetes can impact speech due to nerve damage in head and neck region as well as affecting the larynx and vocal cord elasticity, potentially altering voice production (Saghiri et al., 2022; Ulanovsky et al., 2014). Additionally, diabetes can affect respiratory function (Heimer et al., 1990; Kabitz et al., 2008; Fuso et al., 2012; Zineldin et al., 2015; Pieniawska et al., 2012). Diabetic patients, particularly those with T2DM, exhibit shorter phonation durations and increased vocal straining and hoarseness, often correlated with diabetic neuropathy and poor glycemic control (Hamdan et al., 2013; Gölaç et al., 2022).

While most research has focused on the chronic effects of diabetes on voice, studies examining the short-term impacts of fluctuating glucose levels are limited. One study identified different modulation of voice characteristics during hypoglycemia and hyperglycemia in T1DM patients

(Czupryniak et al., 2019), and other research has suggested that voice characteristics may help estimate glucose levels in non-diabetic individuals (Jeon et al., 2020).

Machine learning techniques have been employed to detect voice changes related to blood glucose fluctuations. Early efforts by (Tschope et al., 2015) utilized a linear regression model with speech features from one T2DM patient and one healthy subject. Subsequent studies, such as that by (Pompe et al., 2023), applied eGeMAPS features and SVM classifiers on speech data from diabetic patients. However, there remains a lack of robust data obtained in standardized clinical settings.

In this chapter, we address the need for continuous, convenient blood glucose management in diabetic patients by leveraging machine learning techniques to identify vocal biomarkers associated with hypoglycemia. Our contribution towards this goal is as following:

- Firstly, we collect a novel database including speech and physiological data from T1DM patients in a controlled clinical setting. The data is collected during euglycemic (normal blood glucose levels) and hypoglycemic (low blood glucose levels) states and is annotated with gold-standard blood glucose (BG) levels.
- We develop machine learning methods to classify between euglycemia and hypoglycemia using both knowledge-based acoustic features and data-driven features (see Section 2.1.1.1). We analyse the performance of the classifiers across different speech tasks and examine the inter-speaker variability in the classification.
- We examine the acoustic features that are most effective in distinguishing between euglycemia and hypoglycemia.
- Finally, we investigate the potential of integrating physiological data, such as ECG signals, to enhance hypoglycemia detection. We also examine previously developed methods to estimate heart rate from speech signals in the context of hypoglycemia detection.

The rest of the chapter is organized as follows. We present our study design in Section 7.2. The data collection procedure and the protocols used, as well as the input features and classification methods, are detailed in Section 7.3. We present the results of the classification experiments in Section 7.4. Additional analysis is provided in Section 7.5. Finally, we conclude the chapter in Section 7.6. The material presented in this chapter has not yet been published elsewhere.

7.2 Study Design

The research has shown that blood glucose levels can affect speech production. However, the aspects of speech that are most affected by hypoglycemia are not well understood. In

this study, we aim to identify biomarkers in speech that can be indicative of hypoglycemia. Towards this goal, we investigate the utility of an extensive set of speech features, previously used in paralinguistic studies. These features are used with a binary classifier to distinguish between euglycemia and hypoglycemia state as shown in Figure 7.1.

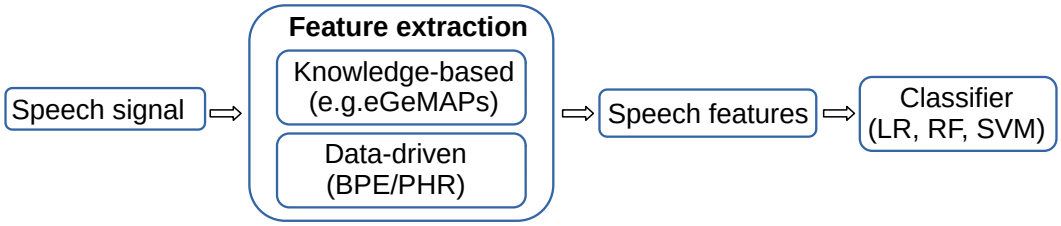


Figure 7.1 – Proposed pipeline for detecting hypoglycemia using various speech features. BPE stands for breathing pattern estimation and PHR stands for phoneme recognition.

The speech features that we utilize in this study are as follows:

- **Knowledge-based:** We extract three sets of knowledge-based features, namely eGeMAPS, ComParE, and long-term average spectrum (LTAS). eGeMAPS and ComParE are widely-used features in paralinguistic and emotion studies in speech (see Section 2.1.1.1). We have also used these features throughout this thesis for various tasks (see Section 4.2.1 and Chapter 6). eGeMAPS has been used to study the voice changes due to fluctuations in blood glucose level (Pompe et al., 2023). LTAS provides spectral information of signal over extended period of time. They have been used in clinical studies of voice quality and pathological speech (see Section 2.1.1.1).

- **Data-driven:** In Chapter 4 we demonstrated the utility of embeddings extracted from pre-trained neural networks for auxiliary tasks. Following the same approach, we use embeddings extracted from raw waveform based CNNs trained for two different tasks, (a) breathing pattern estimation (BPE) and (b) phoneme recognition (PHR). Embeddings extracted from these networks have been used previously in Chapter 4 for various speech related tasks.

For analysis, we further explore the effect of using physiological data, such as ECG signals, to enhance the detection of hypoglycemia. Additionally, we examine the methodology we developed in Chapter 6 to estimate heart rate from speech signals in different glycemic states.

7.3 Experimental Setup

In this section, we first describe the data collection procedure. The speech features extracted and the experimental protocol is followed by the classification models used in our experiments.

7.3.1 Data collection

The database referred to as HypoVoice was collected between November 2022 and February 2023, at the University Clinic for Diabetology, Endocrinology, Nutritional Medicine and Metabolism (UDEM), Insel Hospital, Bern, Switzerland. Ethical approval for the study was granted by the Cantonal Ethics Committee of Bern (KEK-BE: 2022-01142). The database includes simultaneous recordings of speech and physiological signals from 6 patients with type 1 diabetes (3 males and 3 females), in a controlled clinical environment. Data collection was performed under both euglycemic (normal blood glucose levels) and hypoglycemic (low blood glucose levels) states and is annotated with gold-standard BG levels. In the following, we describe the data collection procedure.

7.3.1.1 Inclusion and exclusion criteria

We asked individuals with type 1 diabetes to participate in our study. They were native speakers in German or Swiss German, aged between 21 and 60 years, with an $HbA_{1c} \leq 9.0\%$, and using a continuous glucose monitoring (CGM) system with either multiple daily injections, insulin pump, or hybrid closed-loop insulin therapy. Key exclusion criteria included contradictions to the insulin used to induce the controlled hypoglycemic state; pregnancy or breastfeeding; severe organ dysfunction; cardiovascular or cardiac disease; epilepsy or seizure disorders; drug or alcohol abuse; chronic neurological or ear-nose-and-throat (ENT) disease influencing voice; history of voice disorder; illiteracy; dyslexia; active smoking; or medication known to interfere with voice. 6 individuals were selected to participate in the study.

7.3.1.2 Controlled glycemic states

We adapted the study protocol established in (Lehmann et al., 2024) to introduce different glycemic states (see Figure 7.2). Blood glucose levels were carefully monitored through frequent measurements and regulated using the administration of insulin or glucose.

Participants were required to attend two visits at the Inselspital clinical research unit. During the first visit, informed consent was obtained, physical examinations were conducted, and participants were assessed for eligibility. A preliminary recording session (denoted as m1) was conducted to allow participants to familiarize themselves with the data collection procedure. For the main visit, participants were admitted to the clinical research unit in the early morning following an overnight fast. Throughout the experimental procedure, two medical personnel, including at least one physician, monitored each participant. Before starting the experiment, participants completed an additional test recording session (denoted as m2 in Figure 7.2) to mitigate potential learning effects. Subsequently, their BG level was stabilized in euglycemic range (BG 5.0–8.0 mmol/L), followed by three recordings (sessions m3-m5). Insulin was then administered to induce a hypoglycemic state (BG 2.0–3.5 mmol/L), during which three recordings were collected (sessions m6-m8). Finally, oral glucose was provided to restore BG

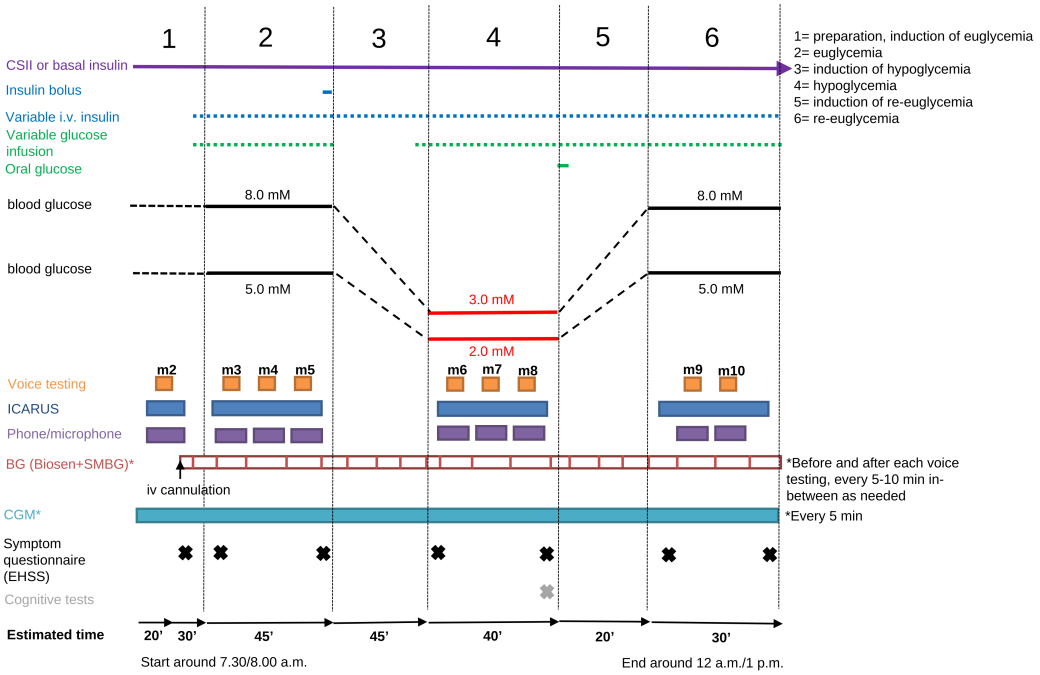


Figure 7.2 – Data collection procedure for collecting simultaneous recordings of speech and physiological signals in euglycemic and hyperglycemic states. The procedure has originally been introduced in (Lehmann et al., 2024).

levels back to the euglycemic range (re-euglycemia), and two recordings were taken during this phase (sessions m9-m10). During this period (re-euglycemia), BG levels were not maintained within a specific range. Each recording lasted 5 to 7 minutes, interspersed with breaks of 3 to 5 minutes for intermittent BG measurements. The procedural details are illustrated in Figure 7.2.

7.3.1.3 Recording sessions

In each session (m1-m10), simultaneous recordings of speech and physiological signals were collected using a smartphone and the ICARUS device (see Section 6.3.1.1), respectively. Similar to the recording setup for the TIPS database in Section 6.3.1.1 the smartphone was placed approximately 20 cm from the participant's mouth and the ICARUS device was worn around the chest. Among the physiological signals recorded, we focused on ECG signals for this study. The recorded speech from the phone and the ECG signals from the ICARUS device are aligned and segmented based on the tasks performed by the participants in each session. The speech tasks performed by the participants are as follows:

- **Sustained vowel:** Participants were requested to vocalize a vowel continuously for approximately 3 seconds, specifically the vowels /a/, /i/, and /u/.
- **Read speech:** Participants read a passage in Swiss German, with each passage being

Table 7.1 – Number of audio files from the HypoVoice dataset used in this study.

Number of audio recordings		Tasks			
		Sustained vowel	Read speech	DDK	Picture description
For each session		3	1	2	1
Total number of euglycemia	For each participant	9	3	6	3
	For all the participants	54	18	36	18
Total number of hypoglycemia	Each participant	9	3	6	3
	All the participants	54	18	36	18

approximately 255 words and assigned randomly.

- **DDK:** Participants executed a Diadochokinetic (DDK) task, repeating two sets of vowels and consonants repeatedly for around 7 seconds each. The first set was /pa/→/ta/→/ka/, and the second set was /ba/→/da/→/ga/.
- **Picture description:** Participants were tasked with describing a displayed picture in Swiss German for around one minute.

Data were collected from 6 participants, who were assigned random IDs ranging from 102 to 107. Given that BG levels were only fully stabilized during the euglycemic and hypoglycemic phases of the study, data collected during these sessions (m3-m8) were utilized for the analyses conducted in this chapter.

Table 7.1 summarizes the number of audio files included in the experimental analyses.

7.3.2 Speech feature extraction

We use 3 sets of knowledge-based features and 3 sets of data-driven features for this study.

Knowledge-based: The 88 dimensional feature set eGeMAPS and the 6373 dimensional feature set ComParE are extracted using the openSMILE toolkit. The LTAS is obtained by calculating a 1024-point discrete Fourier transform of audio frames with a window length of 250 ms and hop size of 10 ms. The log spectrum is calculated for the positive frequencies for each frame. The utterance-level representation is obtained by averaging and calculating the standard deviation of the log spectrum over the duration of the utterance. The resulting 1024 dimensional vector is used as the LATS feature.

Data-driven: We use the raw waveform based CNNs pre-trained for breathing pattern estimation (BPE) and phoneme recognition (PHR), previously used in Chapter 4 for speech related tasks. The utterance-level embeddings are obtained same as before by averaging and calculating the standard deviation of the embeddings over the duration of the utterance. For the BPE networks, we use two models trained with 3 seconds of speech signal as input using mean squared error loss function. One is trained on Philips database and the other on UCL-SBM database. The 20-dimensional embeddings extracted from these networks are denoted as

CNN_Philips and CNN_UCL respectively. The 2028-dimensional embeddings extracted from the PHR network is denoted as CNN_AMI as this network is trained on AMI dataset (Carletta, 2007).

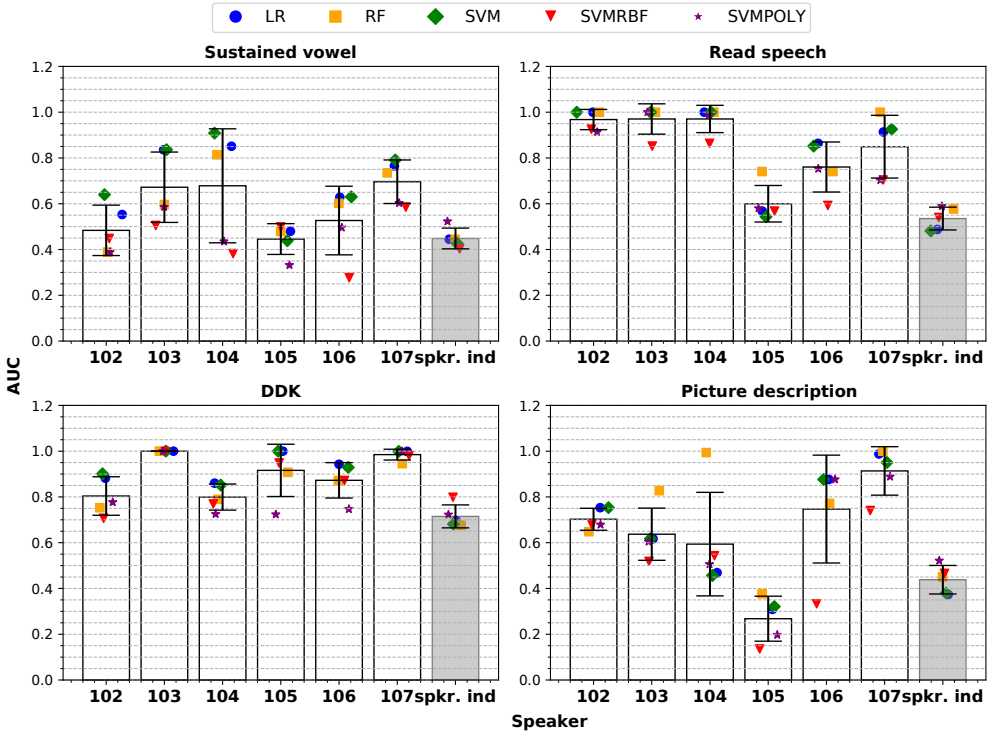


Figure 7.3 – AUC of different classifiers for the speaker independent and speaker specific protocols for four speech tasks. The input feature set is the eGeMAPS.

7.3.3 Experimental protocols

This study is designed to investigate the effect of hypoglycemia on speech in diabetic patients. Many factors could influence the speech characteristics of the participants, such as the duration of diabetes, the severity of the disease, and the medication used to manage it, as well as the individual's general health. We also observed that the performance of systems designed to predict heart rate from speech signals is highly speaker dependent. Therefore, in this study, we also investigate the speaker dependency of the classifiers designed to detect hypoglycemia from speech signals. We repeat the experiments for each speech task separately. The protocols used for the experiments are as follows:

- **Speaker independent:** We use data from all speakers and apply 6-fold cross validation. In each fold, data from one speaker is reserved for testing, while the remaining speakers' data is used for training the models. After completing all folds, the predictions are concatenated, and the overall classifier performance is evaluated.
- **Speaker specific:** In this protocol, we train a new model for each speaker. The number of

utterances is much lower in this case. To ensure that the model performance is reliable we employed a cross validation process. For each fold, one positive sample (utterance during hypoglycemic state) and one negative sample (utterance during euglycemic state) are set aside for testing, with the remaining data used for training. The positive and negative samples in the *Test* set are randomly selected with replacement. This procedure results in a different number of folds per task. Specifically, the number of folds for sustained vowel, read speech, DDK, and picture description are 81, 9, 36, and 9, respectively. The performance is reported by concatenating the results from all the folds for each task.

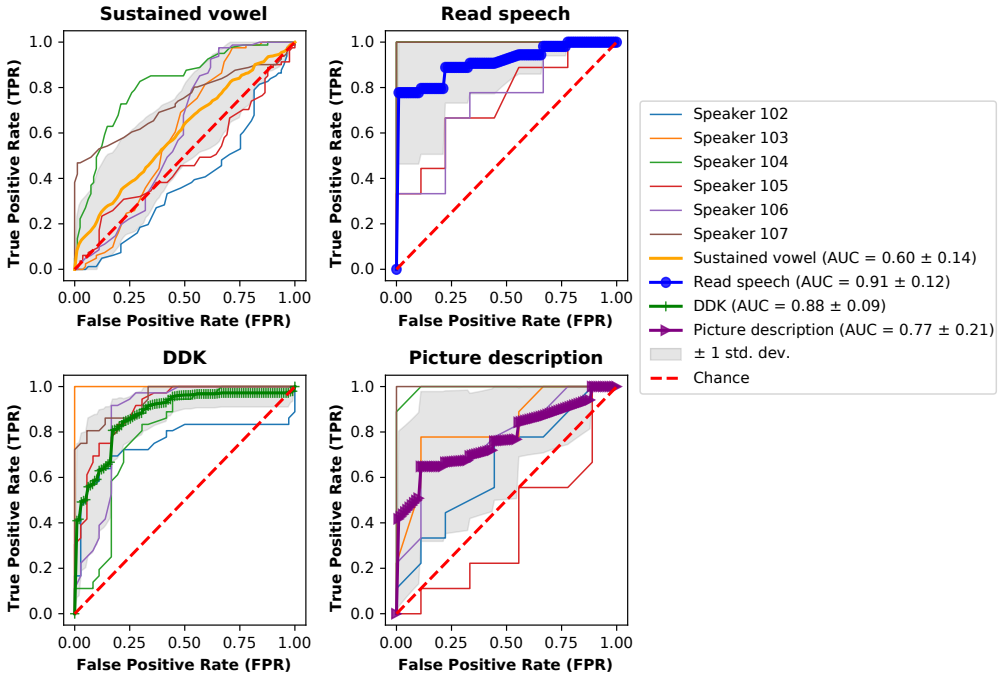


Figure 7.4 – The mean ROC curves over all the speakers for RF classifier. Input features is eGeMAPS.

7.3.4 Classification

Considering the small size of the dataset, we employ classical machine learning approaches for the classification task (see Section 2.1.2.1). Five different classifiers are used in this study to demonstrate the robustness of the proposed feature sets. The classifiers used are as follows: logistic regression (LR), Random Forest (RF), support vector machine with three different kernels; linear (SVM), radial basis function (SVMRBF), and polynomial with degree 3 (SVM-POLY). The default hyperparameters in *Scikit-learn* (Pedregosa et al., 2011) toolkit is used in our experiments without any fine-tuning.

Table 7.2 – The performance metrics for read speech and DDK tasks for all the classifiers when eGeMAPS is used as input features. The reported metrics are AUC, F1 Score (F1), Sensitivity, and Specificity.

Classifier	AUC	F1	Sensitivity	Specificity
Read speech				
LR	0.89±0.17	0.81±0.26	0.82±0.27	0.81±0.25
RF	0.91±0.13	0.88±0.18	0.89±0.17	0.88±0.19
SVM	0.89±0.18	0.80±0.24	0.80±0.24	0.80±0.24
SVMRBF	0.75±0.15	0.64±0.17	0.63±0.17	0.66±0.18
SVMPOLY	0.82±0.17	0.73±0.23	0.72±0.23	0.73±0.22
DDK				
LR	0.95±0.06	0.89±0.12	0.88±0.12	0.89±0.11
RF	0.88±0.10	0.83±0.10	0.84±0.10	0.82±0.11
SVM	0.95±0.06	0.89±0.11	0.90±0.11	0.89±0.11
SVMRBF	0.88±0.12	0.78±0.13	0.77±0.15	0.78±0.12
SVMPOLY	0.83±0.13	0.76±0.19	0.77±0.19	0.76±0.19

7.4 Results

Figure 7.3 illustrates the AUC for all the classifiers across various speech tasks. Both speaker independent and speaker specific protocols are represented in this figure. It can be seen that the performance of individual classifiers varies; however, an overall pattern is emerging. Across all tasks, performance levels differ among speakers, with the speaker independent protocol consistently having lower performance. This observation highlights the speaker dependent nature of speech variations due to hypoglycemia. Furthermore, on average, the read speech and DDK task yield the highest performances. Interestingly, for the speaker independent protocol, the DDK task performs relatively better than the others. The simplicity of the task makes it easier for the participants to perform which could potentially be useful in future applications.

Figure 7.4 shows the mean receiver operating characteristic (ROC) curve (see Section 2.3.1) for the four different tasks, utilizing eGeMAPS as input features and Random Forest (RF) as the classifier. the ROC curve for individual speakers in the read speech and DDK tasks consistently exceeds the chance level (AUC=0.5). Conversely, the sustained vowel task demonstrates the lowest performance among the assessed tasks, aligning with the findings in Figure 7.3, which displays the average performance across all classifiers.

Table 7.2 presents the performance metrics for five classifiers on the read speech and DDK tasks using eGeMAPS as input features. The reported metrics include AUC, F1 Score (F1), Sensitivity, and Specificity (see Section 2.3.1). The RF classifier is the best-performing classifier for the read speech task, while logistic regression (LR) and support vector machine with a linear kernel (SVM) yield the highest performance for the DDK task, with RF following as a

close second based on AUC. For simplicity, the remainder of this chapter will focus on the results from the RF classifier applied to the read speech and DDK tasks.

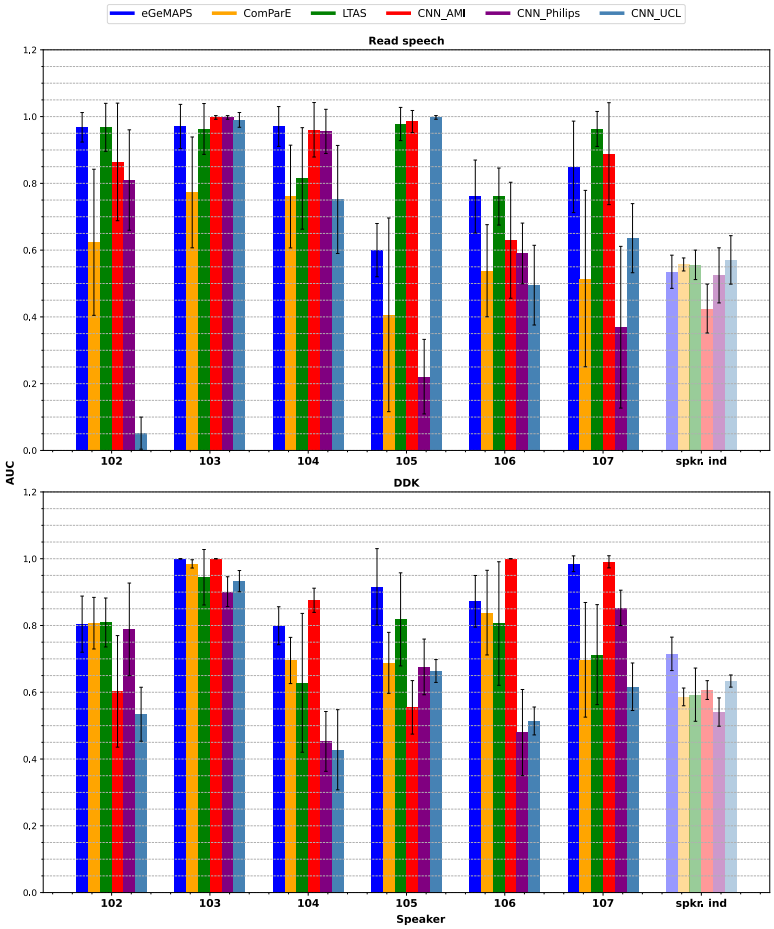


Figure 7.5 – The average AUC from all the classifiers for all the 6 different input features. The performance for both speaker specific and speaker independent protocols is shown.

Similar experiments are conducted using all the previously mentioned features in Section 7.3.2, namely eGeMAPS, ComParE, LTAS, CNN_Philips, CNN_UCL, and CNN_AMI. The performance outcomes indicate a speaker dependent effect, with overall performance being lower in the speaker independent framework for all the above-mentiond features as illustrated in Figure 7.5. Table 7.3 details the RF classifier's performance on the read speech and DDK task for all the input features. For both tasks, the LTAS feature set outperforms the other feature sets. The second best performing feature set is CNN_AMI for read speech task, and eGeMAPS for the DDK task. The embeddings extracted from the BPE networks exhibit the lowest performance.

Table 7.3 – The performance metrics for read speech and DDK tasks for the RF classifier and all the input features. The reported metrics are AUC, F1 Score (F1), Sensitivity, and Specificity.

Feature	AUC	F1	Sensitivity	Specificity
Read speech				
eGeMAPS	0.91±0.13	0.88±0.18	0.89±0.17	0.88±0.19
ComParE	0.94±0.09	0.88±0.15	0.87±0.16	0.89±0.15
LTAS	0.96±0.09	0.92±0.15	0.93±0.13	0.92±0.16
CNN_AMI	0.96±0.11	0.92±0.15	0.91±0.18	0.93±0.12
CNN_Philips	0.70±0.37	0.70±0.31	0.70±0.31	0.70±0.31
CNN_UCL	0.69±0.36	0.67±0.33	0.67±0.33	0.67±0.33
DDK				
eGeMAPS	0.88±0.10	0.83±0.10	0.84±0.10	0.82±0.11
ComParE	0.86±0.13	0.78±0.16	0.78±0.16	0.78±0.16
LTAS	0.88±0.18	0.86±0.16	0.86±0.16	0.86±0.16
CNN_AMI	0.81±0.22	0.79±0.21	0.79±0.21	0.79±0.20
CNN_Philips	0.68±0.21	0.64±0.19	0.64±0.20	0.65±0.19
CNN_UCL	0.65±0.13	0.63±0.12	0.63±0.11	0.64±0.12

7.5 Analysis of Proposed Approaches

In this section, we analyse the features to understand the most important ones in the classification task. We also investigate the potential of integrating ECG information to enhance the detection of hypoglycemia. Finally, we examine the methodology developed in Chapter 6 to estimate heart rate from speech signals in different glycemic states.

7.5.1 Feature importance

A feature ranking analysis is performed on the eGeMAPs features for the RF classifier. Figure 7.6 depicts the categories of the top 10 important features for both the read speech and DDK tasks across all speakers. The frequency with which each feature category appears in the top 10 important features for each speaker is recorded. Each speaker is represented by a different color to indicate the feature categories that frequently appear for them. For the read speech task, features related to source characteristics, such as fundamental frequency (F0) and shimmer, consistently appear across all speakers, alongside system related features like slope. Mel frequency cepstral coefficients (MFCCs) are also indicated for all but one speaker. In contrast, the most crucial feature categories for the DDK task are loudness and MFCCs; however, they do not consistently emerge across all speakers. This might be due to the fact that DDK recordings are much shorter and less complex than the read speech task.

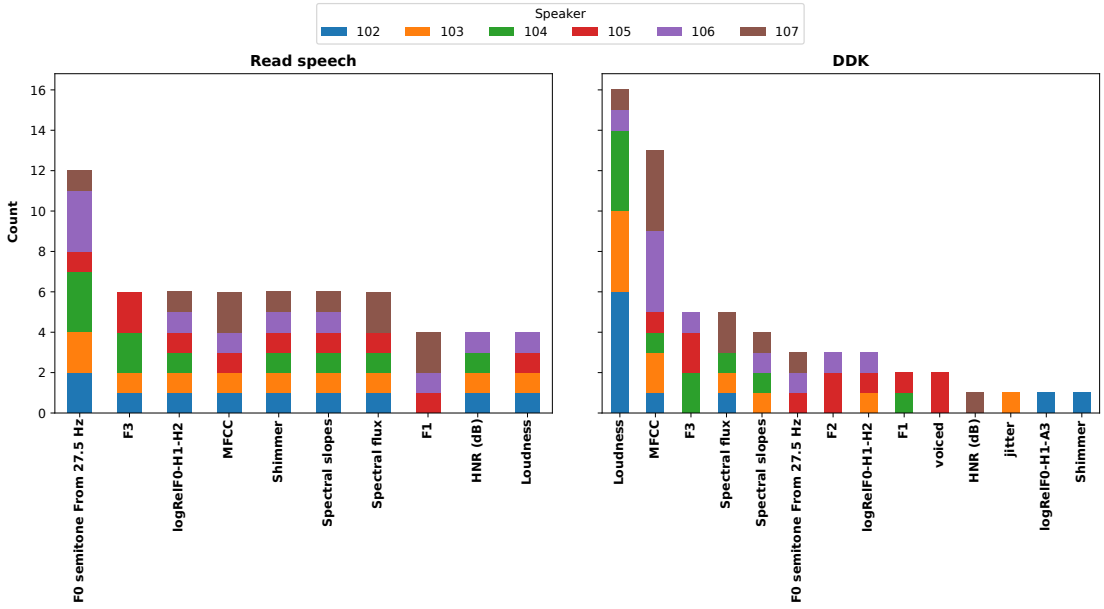


Figure 7.6 – The top 10 important feature categories among all speakers for read speech and DDK tasks

7.5.2 Incorporating cardiac information

Hypoglycemia causes physiological responses within the body, notably affecting the cardiovascular system. The hypovoice database includes ECG signals collected simultaneously with speech signals during euglycemic and hypoglycemic phases. This allows us to investigate their utility in distinguishing between the two phases either independently or in conjunction with speech features.

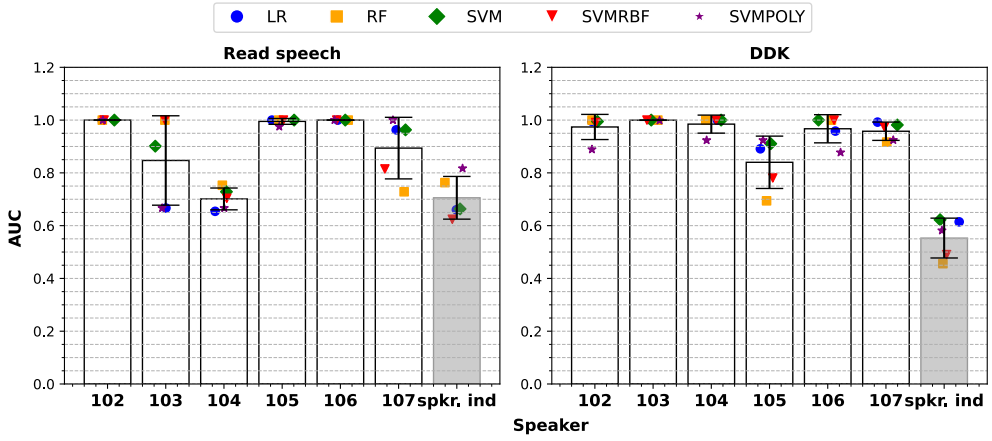


Figure 7.7 – AUC of different classifiers for all the speakers as well as for the speaker independent protocol. The input features is the ECG inter-beat intervals (IBIs).

In this section, we use inter-beat intervals (IBI) extracted from the ECG signals as physiological

features. The IBI represent the intervals between consecutive R peaks in the ECG signal. The algorithm requires a minimum of 1 minute of data to identify R peaks; hence, the original continuous recordings from the ICARUS device are utilized to extract a continuous signal of IBI. This continuous signal (IBI) was provided to us by CSEM, and we did not extract it ourselves. We align the provided IBI signal with the speech signals and segment them based on the tasks and recording sessions. The average and standard deviation of the IBI are calculated and concatenated to form a 2-dimensional feature vector. We use the 2-dimensional IBI feature set in two ways: (a) as an independent feature set input to the classifiers, and (b) combined with the speech features (early fusion). The same speaker independent and speaker specific protocols are applied as before.

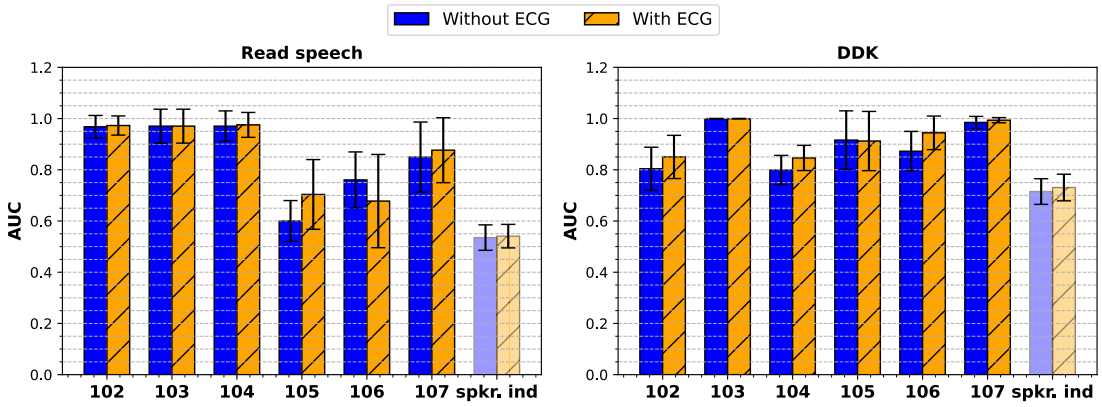


Figure 7.8 – Comparing AUC of different classifiers when ECG detived information (IBI) is fused with speech features or not. The input speech feature is eGeMAPS .

Figure 7.7 presents the AUC values for various classifiers using the 2-dimensional IBI feature as input for read speech and DDK tasks. It is important to note that ECG recordings from one session of subject 106 were corrupted, resulting in the exclusion of these samples from analysis. The results indicate that this 2-dimensional IBI feature set effectively distinguishes between the two states across all tasks. Figure 7.8 presents the classification performance for read speech and DDK tasks when the IBI feature set is concatenated with the eGeMAPS features (early fusion). The results show that the fusion of the IBI feature set enhances the performance of the classifiers.

7.5.3 Analysis of estimated cardiac activity from speech signals

In chapter 6, we introduced a method to estimate heart rate from speech signals. In that study, there was no well-defined physiological states which could affect the heart rate. The HypoVoice database, however, provides ECG signals from participants in two different glycemic states. The glycemic state is affecting the cardiac parameters as presented in Section 7.5.2. This provides an opportunity to investigate the validity of the developed methods in estimating heart rate from speech signals in different physiological states. Previously, we evaluated the performance of the method based on the ground truth heart rate. In this case, not the

similarity between the estimated heart rate and the ground truth, but the difference between the estimated heart rate in euglycemia and hypoglycemia is of interest.

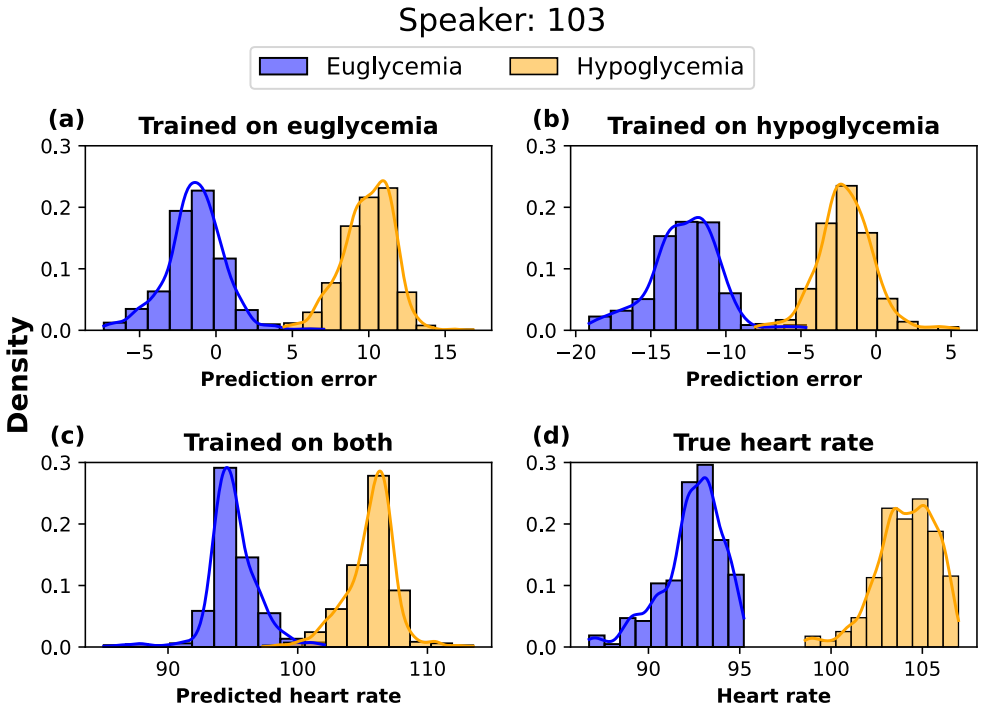


Figure 7.9 – The interquartile range (IQR) of estimated heart rate prediction error for systems trained on only euglycemia or hypoglycemia, as well as the IQR of predicted heart rate from the system trained on both set of data for speaker 103.

We train models on one glycemic state and test it on both glycemic states. We examine the distribution of the error between estimated heart rate and the ground truth for both states. If the error distribution is well separated between the two conditions, it indicates that the estimated heart rate is indicative of the glycemic state.

We employed the Hybrid BYOL-S features used in Section 6.3.3 with input window size of 5 seconds as input features and trained four regression models, namely Ridge regression, Random Forrest (RF), Gradient Boosting tree (GBT), and multi-layer perceptron (MLP) as described in Section 6.3.5. We followed the speaker specific protocol presented in Section 6.3.4 for training our models meaning we train a different model for each speaker. The read speech task is chosen for this analysis.

Two training scenarios are considered: (a) a model trained solely on euglycemic data, evaluated on both euglycemic and hypoglycemic conditions, and (b) a model trained exclusively on hypoglycemic data, evaluated similarly. The training data includes two sessions from each condition, while the evaluation data consists of one session from each condition. We examine the distribution of the error in these scenarios. Additionally, for each speaker, we train a model on data from both euglycemic and hypoglycemic states and evaluate it on the same evaluation

data as before. We visualize the distribution of the estimated heart rate in this case.

Figure 7.9 presents the results for speaker 103. In this case, the ground truth heart rates are well separated between the two conditions which is also reflected in the performance of the previously presented classifier on the IBI features (see Figure 7.7). It can be seen that the error distribution is also well separated between the two conditions. This indicates that the models trained for estimating heart rate from speech signals are indeed able to capture relevant cardiac activity information. When the model is trained on both conditions, it is able to model the distribution of the heart rate in both conditions.

Interestingly, it can be observed that the heart rate estimation models are also behaving in a speaker dependent manner. They are dependent on the ground truth heart rate that is provided to them. For example, for speaker 104, presented in Figure 7.10, there is not a significant difference in the ground truth heart rate between the two conditions, resulting in low classification performance with IBI features (see Figure 7.7). This is also reflected in the distribution of the estimated heart rate error when the models are trained on one condition as well as the estimated heart rate when the model is trained on both conditions.

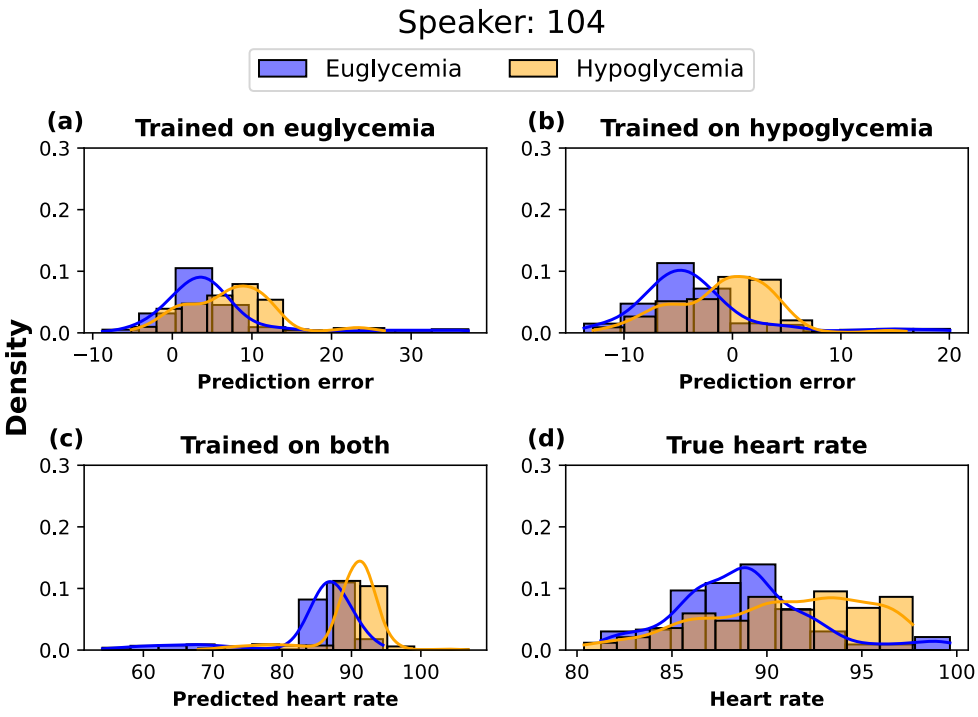


Figure 7.10 – The interquartile range (IQR) of estimated heart rate prediction error for systems trained on only euglycemia or hypoglycemia, as well as the IQR of predicted heart rate from the system trained on both set of data for speaker 104.

These findings can be considered as a fist step for validating our methodology for estimating cardiac activity information from speech signals, emphasizing its potential applications within medical settings.

7.6 Conclusion

In this chapter, we studied the potential of using speech signals to detect hypoglycemia in individuals with type 1 diabetes. We introduced a novel dataset that includes both speech and ECG signals collected from participants in euglycemic and hypoglycemic states in a clinical setting. The participants were asked to perform four speech tasks, namely sustained vowel, read speech, diadochokinetic (DDK), and picture description. A variety of knowledge-based and data-driven features were extracted from the speech signals, and the performance of multiple classifiers was evaluated across four distinct tasks. Additionally, we employed our experiments using a speaker independent and speaker specific protocol. We observed that this task is very speaker dependent. The results showed that read speech and DDK tasks were most effective in differentiating between the two glycemic states based on speech signals. Long-term average spectrum (LTAS) features outperformed other feature sets in both tasks. Additionally, our analysis of feature importance revealed that both source related and system related features play a critical role in classification. Incorporating cardiac information alongside speech features further improved classifier performance. Lastly, we revisited a previously developed method for estimating heart rate from speech signals in different glycemic states. We demonstrated that these models are able to capture the distribution of the heart rate based on the ground truth heart rate they observe during training. For example, if a model is trained only on hypoglycemia data, they are able to model the distribution of the heart rate in hypoglycemia while over predicting the heart rate in euglycemia. Additionally, we observed that there was a speaker dependent effect even in the ground truth heart rate. For some speakers, the heart rate was well separated between the two conditions, while for others, it was not. This speaker dependent effect was also reflected in the estimated heart rate from speech signals. This emphasizes that the models for estimating heart rate from speech signals are also speaker dependent.

8 Conclusions and Future Directions

Speech is a complex signal generated by many physiological systems working in synchrony. The speech production system is capable of conveying a wide range of information, from the semantics of the spoken words to speaker dependent characteristics. Changes in these physiological systems can lead to variations in speech. This thesis studied the relationships between speech and physiological signals such as breathing patterns and heart rate using machine learning techniques and incorporated physiological information into speech related applications.

We developed end-to-end convolutional neural network (CNN) models to estimate breathing patterns from raw waveform speech signals. We based our models on a CNN architecture that had previously been used for classification tasks. By adapting this architecture for estimating breathing patterns, we demonstrated its suitability for regression tasks as well. We compared our approach to a spectral based method developed in parallel which used Log Mel spectograms instead of raw waveforms as input. Model performance was assessed using breathing related parameters, including breathing rate, breath event sensitivity, and tidal volume in addition to common regression metrics such as mean squared error (MSE) and Pearson's correlation coefficient. Our study demonstrated that both approaches were performing similarly. However, the spectral based method required a minimum of 4 seconds of speech for accurate breathing pattern estimation, while the raw waveform based models achieved similar performance with smaller segments of speech even as short as 2 seconds. We evaluated our models in a cross database setting, and observed that even though the performance dropped in terms of Pearson's correlation, specially in raw waveform based approach, the models were still able to estimate breathing parameters reasonably well. Additionally, we investigated the limitations of the metrics used for evaluating the performance of the models and demonstrated that the commonly used regression evaluation metrics such as MSE and Pearson's correlation may not adequately reflect the performance of breathing pattern estimation models. This was further corroborated by (Deshpande et al., 2023).

Furthermore, we investigated the information modeled by the raw waveform based networks by systematically limiting the frequency content of the input speech signals and analysing

the impact on network outputs, extracted neural embeddings, and the model performance in two downstream tasks. Our analysis indicated that raw waveform based breathing pattern estimation networks relied on low frequency regions of speech signals when the input window was 3 seconds and specially the frequencies lower than 500 Hz were very informative for these models. When the input window was reduced to 2 seconds, the networks started also relying on higher frequency regions of the speech signals. This suggests that the raw waveform based networks are modeling phonetic level acoustic information as well as source related information such as pitch. Voiced consonants and vowels have higher energy in lower frequency regions. These events, which are produced by the vibration of the vocal folds, seem to be important for the raw waveform based models for breathing pattern estimation. This understanding might also help explain why spectral based methods required longer input windows compared to raw waveform based models. In the spectral based approach, the network is modeling the envelope of the spectrum which predominantly captures the formant information. They do not focus much on the pitch information whereas in the raw waveform based models the network is capturing both the formant and pitch information. Therefore, the spectral based models require longer input windows to capture enough formant information to estimate the breathing patterns accurately. This study has practical implications as well. For example in telehealth applications the speech signal is transmitted over a low bandwidth channel and hence has a limited frequency content. Other applications include privacy-preserving breathing pattern estimation, and design of efficient wearable sensors for speech-based breathing pattern estimation.

We extracted neural embeddings from the layer before the output of the raw waveform based models (10 dimension) and used them in various speech based tasks, including COVID-19 detection, distinguishing between natural and synthetic speech, and emotion recognition. We compared the performance of these embeddings with traditional knowledge-based features and neural embeddings from networks pre-trained for other tasks. In COVID-19 detection, performance was comparable to the knowledge-based features (around 2% lower AUC on the *Test* set with 7% higher sensitivity) but lower than pre-trained embeddings for phoneme recognition (7% lower AUC and 15% lower sensitivity). A reason for this could be that the respiratory issues caused by COVID-19 infection in the database were not severe enough to be detected by the models. The changes in the speech was more captured in the phoneme level embeddings than the breathing pattern. This observation is corroborated with findings in (Deshpande et al., 2021a) in which they showed that COVID-19 infection has a higher impact on the properties of vocal tract modulation than the source of excitation. For the second task, we investigated the difference between breathing pattern information in natural and synthetic speech. We used a presentation attack detection framework in this study in which we showed that our models effectively differentiate between natural and synthetic speech produced by text-to-speech (TTS) systems with the highest AUC of 99.93%. It should be noted that these performances were achieved using a very low dimensional neural embeddings (20 dimension using mean and standard deviation to generate utterance-level features) and a very simple classifier (Random Forest). They completely failed to distinguish between natural

and synthetic speech when voice conversion (VC) methods were used to generate synthetic speech. In VC methods, a person's real speech signal is manipulated to sound like a target speaker, whereas TTS methods generate speech without relying on a real speech signal. This suggests that the breathing pattern information captured by our networks is still present even after the voice manipulation in VC methods. For emotion recognition, our neural embeddings improved the baseline models trained on the ground truth physiological signals by 30% but were outperformed by around 60% by networks pre-trained specifically for emotion and phoneme recognition. This suggests that the physiological information alone are not capable of capturing the emotional state of the speaker.

We developed models to estimate cardiac activity parameters, specifically heart rate, from speech signals using knowledge-based features (eGeMAPS and ComParE) and neural embeddings extracted from a self-supervised model (Hybrid BYOL-S). A novel database was created, comprising simultaneous recordings of speech and physiological signals collected while participants performed two tasks: read speech and free speech. This dataset, along with another database originally designed for emotion recognition research, was utilized in our studies. Results indicated that across all three datasets, models leveraging neural embeddings consistently outperformed those based on knowledge-based features, with performance improving as the context window increased. On average, we observed an 8% improvement in accuracy when extending input speech from 3 to 5 seconds using neural embeddings, with an even greater improvement when using knowledge-based features. We found that the methodology is highly speaker dependent and subject to intra-speaker variability. Heart rate variation arises from multiple factors, including physiological, emotional, and environmental influences, complicating the development of speaker independent models. Nevertheless, this approach shows promise for medical applications, as it can be personalized for long-term monitoring of individuals.

Finally, we studied a clinical application aimed at identifying vocal biomarkers of hypoglycemic state in diabetic patients. We collected a novel database of simultaneous recordings of speech and physiological signals during euglycemic and hypoglycemic states in a controlled clinical environment. We employed neural embeddings from previously used pre-trained models and knowledge-based features to detect hypoglycemia in diabetic patients across various speech tasks. Our findings revealed that the long-term average spectrum (LTAS) is a reliable indicator for hypoglycemia detection and that incorporating physiological information, such as the inter-beat interval (IBI), can enhance model performance. Additionally, we used the developed models for estimating heart rate from speech signals using the speech data from the two conditions. We observed that the distribution of the estimated heart rates differed between the hypoglycemic and euglycemic states. The clear distinction was directly linked to how the ground truth heart rate were different between the two states. This suggests that the estimated heart rate from speech signals can potentially be used in a medical application such as hypoglycemia detection.

Despite increasing interest in the intersection of speech and physiological signals, considerable

work remains. Following are possible future directions:

Limitations of evaluation metrics and robustness to noise: We employed mean squared error (MSE) and Pearson's correlation coefficient as metrics for training and evaluating models aimed at estimating breathing patterns from speech. However, we have observed that these evaluation metrics do not accurately reflect the characteristics of the breathing parameters. For instance, a high MSE does not necessarily correlate with a significant error in the estimated breathing rate. Integrating breathing parameters into the training process of neural networks can enhance the model's ability to encapsulate relevant physiological information. Additionally, for models developed to estimate physiological signals from speech to be viable in real-world applications, their robustness to noise must be established. An investigation into the impact of noise on model performance is essential.

Leveraging foundation models in estimation of physiological information: The advent of foundation models has significantly transformed the domain of machine learning. These models have demonstrated promising results across various tasks, particularly in paralinguistic studies and pathological speech processing, which often rely on small datasets. We utilized a self-supervised model to estimate heart rate from speech signals. Recently, (Mitra et al., 2024) employed representations from a foundation model to estimate breathing rate from speech signals . Nonetheless, the application of foundation models in physiological signal processing remains largely unexplored.

More in-depth analysis of models when the ground truth physiological signals are available: Pre-trained models for estimating breathing patterns from speech signals were evaluated by restricting the frequency content of the input speech. This methodological approach was developed around auxiliary tasks in the absence of ground truth physiological signals. When ground truth physiological signals are available, a more in-depth analysis of models could be conducted, for example, through the use of relevance maps (Muckenhirn et al., 2019).

Further exploration of heart rate estimation from speech signals: In the last chapter, we demonstrated that the distribution of the estimated heart rates produced by previously developed models differs between hypoglycemic and euglycemic states. A natural extension of this research is to use the estimated heart rate within a classification framework. Furthermore, their impact when they are incorporated with speech features could be investigated. Additionally, their ability to differentiate between two glycemic states shows their potential, and they could be explored to replace current methods.

Incorporating speech and physiological signals in a multi-modal framework: Early fusion of speech and physiological signals improved system performance for emotion recognition and hypoglycemia detection tasks. This demonstrates the potential of multi-modal frameworks in advancing speech related applications by leveraging complementary information from diverse data sources. Further exploration of fusion strategies is required to better understand their impact on system performance. Additionally, multi-modal representation learning could be employed to further optimize the integration of speech and physiological signals.

Bibliography

Abadi, Martín et al. (Mar. 2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. arXiv:1603.04467 [cs]. DOI: 10.48550/arXiv.1603.04467 (cit. on p. 26).

Amiriparian, Shahin et al. (2017). "Snore sound classification using image-based deep spectrum features". In: *Interspeech 2017*, pp. 3512–3516. DOI: 10.21437/Interspeech.2017-434 (cit. on p. 67).

ASVspoof (2022). URL: https://www.asvspoof.org/ (visited on 2022) (cit. on p. 57).

Avila, Flavio et al. (2021). "Investigating feature selection and explainability for COVID-19 diagnostics from cough sounds". In: *Interspeech 2021*, pp. 951–955. DOI: 10.21437/Interspeech.2021-2197 (cit. on p. 49).

Baird, Alice, Shahin Amiriparian, Miriam Berschneider, Maximilian Schmitt, and Björn W Schuller (2019). "Predicting Biological Signals from Speech: Introducing a Novel Multimodal Dataset and Results". In: *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–5. DOI: 10.1109/MMSP.2019.8901758 (cit. on p. 63).

Baird, Alice, Lukas Stappen, Lukas Christ, Lea Schumann, Eva-Maria Messner, and Björn W Schuller (2021). "A Physiologically-Adapted Gold Standard for Arousal during Stress". In: *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. MuSe '21. Virtual Event, China: Association for Computing Machinery, pp. 69–73. DOI: 10.1145/3475957.3484446 (cit. on p. 63).

Baird, Alice, Andreas Triantafyllopoulos, et al. (Dec. 2021). "An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress". English. In: *Frontiers in Computer Science* 3. Publisher: Frontiers. DOI: 10.3389/fcomp.2021.750284 (cit. on p. 10).

Baird, Alice et al. (July 2022). *The ICML 2022 Expressive Vocalizations Workshop and Competition: Recognizing, Generating, and Personalizing Vocal Bursts*. arXiv:2205.01780 [eess]. DOI: 10.48550/arXiv.2205.01780 (cit. on p. 69).

Barche, Purva, Krishna Gurugubelli, and Anil Kumar Vuppala (2020). "Towards Automatic Assessment of Voice Disorders: A Clinical Approach". In: *Interspeech 2020*, pp. 2537–2541. DOI: 10.21437/Interspeech.2020-2160 (cit. on p. 10).

Bedi, Gillinder et al. (2015). "Automated analysis of free speech predicts psychosis onset in high-risk youths". In: *npj Schizophrenia* 1.1. Publisher: Nature Publishing Group, pp. 1–7 (cit. on p. 1).

Bocklet, Tobias, Andreas Maier, Josef G. Bauer, Felix Burkhardt, and Elmar Noth (2008). "Age and gender recognition for telephone applications based on GMM supervectors and sup-

port vector machines". In: *ICASSP 2008 - 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1605–1608. DOI: 10.1109/ICASSP.2008.4517932 (cit. on pp. 1, 65).

Boiten, Frans A. (1998). "The effects of emotional behaviour on components of the respiratory cycle". In: *Biological Psychology* 49.1, pp. 29–51. DOI: https://doi.org/10.1016/S0301-0511(98)00025-8 (cit. on p. 65).

Boiten, Frans A., Nico H. Frijda, and Cornelis J.E. Wientjes (1994). "Emotions and respiratory patterns: review and critical analysis". In: *International Journal of Psychophysiology* 17.2, pp. 103–128. DOI: https://doi.org/10.1016/0167-8760(94)90027-2 (cit. on p. 63).

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik (July 1992). "A training algorithm for optimal margin classifiers". In: *Proceedings of the 5th Annual Workshop on Computational Learning Theory*. COLT '92. New York, NY, USA: Association for Computing Machinery, pp. 144–152. DOI: 10.1145/130385.130401 (cit. on p. 12).

Bottou, Léon (2010). "Large-Scale Machine Learning with Stochastic Gradient Descent". en. In: *Proceedings of COMPSTAT'2010*. Ed. by Yves Lechevallier and Gilbert Saporta. Heidelberg: Physica-Verlag HD, pp. 177–186. DOI: 10.1007/978-3-7908-2604-3_16 (cit. on p. 12).

Breiman, Leo (Oct. 2001). "Random Forests". en. In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324 (cit. on p. 12).

Brown, Chloë et al. (Aug. 2020). "Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data". en. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Virtual Event CA USA: Association for Computing Machinery, pp. 3474–3484. DOI: 10.1145/3394486.3412865 (cit. on p. 48).

Busso, Carlos et al. (2008). "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language Resources and Evaluation* 42.4. Publisher: Springer, pp. 335–359 (cit. on pp. 67, 69).

Carletta, Jean (2007). "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus". In: *Language Resources and Evaluation* 41.2. Publisher: Springer, pp. 181–190 (cit. on pp. 51, 101).

Christ, Lukas et al. (Oct. 2022). "The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress". In: *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. MuSe' 22. New York, NY, USA: Association for Computing Machinery, pp. 5–14. DOI: 10.1145/3551876.3554817 (cit. on pp. 2, 66, 68, 70, 71, 89).

Collobert, Ronan, Christian Puhrsch, and Gabriel Synnaeve (2016). *Wav2Letter: an End-to-End ConvNet-based Speech Recognition System*. arXiv:1609.03193 [cs.LG]. DOI: 10.48550/arXiv.1603.04467 (cit. on p. 65).

Conrad, B. and P. Schönle (1979). "Speech and respiration". In: *Archiv für Psychiatrie und Nervenkrankheiten* 226.4. Publisher: Springer, pp. 251–268 (cit. on p. 1).

Cortes, Corinna and Vladimir Vapnik (Sept. 1995). "Support-vector networks". en. In: *Machine Learning* 20.3, pp. 273–297. DOI: 10.1007/BF00994018 (cit. on p. 12).

Cummins, Nicholas, Alice Baird, and Björn W Schuller (Dec. 2018). "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning". en. In: *Methods (San Diego, Calif.)* 151, pp. 41–54. DOI: 10.1016/j.ymeth.2018.07.007 (cit. on p. 19).

Cummins, Nicholas, Maximilian Schmitt, Shahin Amiriparian, Jarek Krajewski, and Björn W Schuller (2017). ""You sound ill, take the day off": Automatic recognition of speech affected by upper respiratory tract infection". In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3806–3809. DOI: 10.1109/EMBC.2017.8037686 (cit. on p. 19).

Czupryniak, Leszek et al. (June 2019). "378-P: Human Voice Is Modulated by Hypoglycemia and Hyperglycemia in Type 1 Diabetes". In: *Diabetes* 68.Supplement_1, 378–P. DOI: 10.2337/db19-378-P (cit. on p. 96).

Das, Rohan Kumar, Maulik Madhavi, and Haizhou Li (2021). "Diagnosis of COVID-19 using auditory acoustic cues". In: *Interspeech 2021*, pp. 921–925. DOI: 10.21437/Interspeech.2021-497 (cit. on p. 48).

Deshpande, Gauri and Björn W Schuller (Nov. 2021a). "COVID-19 Biomarkers in Speech: On Source and Filter Components". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. ISSN: 2694-0604, pp. 800–803. DOI: 10.1109/EMBC46164.2021.9629831 (cit. on p. 112).

Deshpande, Gauri and Björn W Schuller (2021b). "The DiCOVA 2021 challenge — An encoder-decoder approach for COVID-19 recognition from coughing audio". In: *Interspeech 2021*, pp. 931–935. DOI: 10.21437/Interspeech.2021-811 (cit. on pp. 48, 50).

Deshpande, Gauri, Björn W Schuller, Pallavi Deshpande, Anuradha Rajiv Joshi, S. K. Oza, and Sachin Patel (2023). "Analysing Breathing Patterns in Reading and Spontaneous Speech". en. In: *Speech and Computer*. Ed. by Alexey Karpov, K. Samudravijaya, K. T. Deepak, Rajesh M. Hegde, Shyam S. Agrawal, and S. R. Mahadeva Prasanna. Cham: Springer Nature Switzerland, pp. 3–17. DOI: 10.1007/978-3-031-48312-7_1 (cit. on p. 111).

Dhall, Abhinav, Garima Sharma, Roland Goecke, and Tom Gedeon (2020). "EmotiW 2020: Driver Gaze, Group Emotion, Student Engagement and Physiological Signal based Challenges". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. ICMI '20. Virtual Event, Netherlands: ACM, pp. 784–789. DOI: 10.1145/3382507.3417973 (cit. on p. 63).

Dibazar, Alireza A, S Narayanan, and Theodore W Berger (2002). "Feature analysis for automatic detection of pathological speech". In: *Proceedings of the 2nd Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*. Vol. 1. IEEE, pp. 182–183 (cit. on p. 19).

Dubagunta, S. Pavankumar, Bogdan Vlasenko, and Mathew Magimai.-Doss (2019). "Learning voice source related information for depression detection". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6525–6529. DOI: 10.1109/ICASSP.2019.8683498 (cit. on pp. 13, 65).

Duffy, J. R. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management*. St. Louis, MO: Elsevier (cit. on p. 56).

El Ayadi, Moataz, Mohamed S Kamel, and Fakhri Karray (2011). "Survey on speech emotion recognition: Features, classification schemes, and databases". In: *Pattern Recognition* 44.3. Publisher: Elsevier, pp. 572–587 (cit. on pp. 1, 65).

Elbanna, Gasser, Alice Biryukov, et al. (2022). "Hybrid handcrafted and learnable audio representation for analysis of speech under cognitive and physical load". In: *Interspeech 2022*, pp. 386–390. DOI: 10.21437/Interspeech.2022-10498 (cit. on pp. 11, 87, 89, 90).

Elbanna, Gasser, Zohreh Mostaani, and Mathew Magimai.-Doss (Sept. 2024). "Predicting Heart Activity from Speech using Data-driven and Knowledge-based features". en. In: *Interspeech 2024*, pp. 4758–4762. DOI: 10.21437/Interspeech.2024-2150 (cit. on p. 86).

Elbanna, Gasser, Neil Scheidwasser-Clow, Mikolaj Kegler, Pierre Beckmann, Karl El Hajal, and Milos Cernak (2022). "Byol-s: Learning self-supervised speech representations by bootstrapping". In: *HEAR: Holistic Evaluation Of Audio Representations*. PMLR, pp. 25–47 (cit. on pp. 87, 90).

Enderby, P. (1980). "Frenchay dysarthria assessment". In: *British Journal of Disorders of Communication* 15.3, pp. 165–173 (cit. on p. 56).

Eyben, Florian, Felix Weninger, Florian Gross, and Björn W Schuller (Oct. 2013). "Recent developments in openSMILE, the munich open-source multimedia feature extractor". In: *Proceedings of the 21st ACM International Conference On Multimedia*. MM '13. New York, NY, USA: Association for Computing Machinery, pp. 835–838. DOI: 10.1145/2502081.2502224 (cit. on p. 9).

Eyben, Florian, Martin Wöllmer, and Björn W Schuller (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459–1462 (cit. on pp. 52, 89).

Eyben, Florian et al. (Apr. 2016). "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". en. In: *IEEE Transactions on Affective Computing* 7.2, pp. 190–202. DOI: 10.1109/TAFFC.2015.2457417 (cit. on pp. 9, 89).

Fairbanks, G (1960). "The rainbow passage". In: *Voice and Articulation Drillbook* 2. Publisher: Harper & Row (cit. on p. 29).

Freund, Yoav et al. (1996). "Experiments with a new boosting algorithm". In: *ICML*. Vol. 96. Citeseer, pp. 148–156 (cit. on pp. xiii, 12, 49, 52).

Fu, Szu-Wei, Yu Tsao, Xugang Lu, and Hisashi Kawai (2017). "Raw waveform-based speech enhancement by fully convolutional networks". In: *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 006–012. DOI: 10.1109/APSIPA.2017.8281993 (cit. on p. 23).

Fuchs, Susanne, Uwe D. Reichel, and Amelie Rochet-Capellan (2015). "Changes in speech and breathing rate while speaking and biking". In: *ICPhS 2015: 18th International Congress of Phonetic Sciences* (cit. on p. 23).

Fuso, Leonello et al. (2012). "Reduced respiratory muscle strength and endurance in type 2 diabetes mellitus". en. In: *Diabetes/Metabolism Research and Reviews* 28.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/dmrr.2284, pp. 370–375. DOI: 10.1002/dmrr.2284 (cit. on p. 95).

Gallardo-Antolín, Ascensión and Juan M. Montero (Oct. 2021). "On combining acoustic and modulation spectrograms in an attention LSTM-based system for speech intelligibility level classification". In: *Neurocomputing* 456, pp. 49–60. DOI: 10.1016/j.neucom.2021.05.065 (cit. on p. 10).

Gölaç, Hakan, Güzide Atalik, Alper Kutalmış Türkcan, and Metin Yilmaz (Oct. 2022). "Disease related changes in vocal parameters of patients with type 2 diabetes mellitus". eng. In: *Logopedics, Phoniatrics, Vocology* 47.3, pp. 202–208. DOI: 10.1080/14015439.2021.1917653 (cit. on p. 95).

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press (cit. on pp. 12, 24).

Griffin, D. and Jae Lim (1984). "Signal estimation from modified short-time Fourier transform". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.2, pp. 236–243. DOI: 10.1109/TASSP.1984.1164317 (cit. on p. 58).

Grimm, Michael and Kristian Kroschel (2005). "Evaluation of natural emotions using self assessment manikins". In: *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005*. IEEE, pp. 381–385 (cit. on p. 63).

Haider, Fasih, Sofia De La Fuente Garcia, Pierre Albert, and Saturnino Luz (2020). "Affective speech for Alzheimer's dementia recognition". In: *LREC: Resources and Processing of Linguistic, Para-Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric/Developmental Impairments (RaPID)*. Publisher: European Language Resources Association (ELRA), pp. 67–73 (cit. on p. 10).

Halpern, Bence Mark, Julian Fritsch, Enno Hermann, Rob van Son, Odette Scharenborg, and Mathew Magimai-Doss (2021). "An Objective Evaluation Framework for Pathological Speech Synthesis". In: *Speech Communication; 14th ITG Conference*, pp. 1–5 (cit. on pp. 10, 56).

Hamdan, Abdul-latif, Jad Jabbour, Randa Barazi, Zeina Korban, and Sami T. Azar (July 2013). "Prevalence of Laryngopharyngeal Reflux Disease in Patients With Diabetes Mellitus". In: *Journal of Voice* 27.4, pp. 495–499. DOI: 10.1016/j.jvoice.2012.07.010 (cit. on p. 95).

Hammarsten, Jonna, Roxanne Harris, Nilla Henriksson, Isabelle Pano, Mattias Heldner, and Marcin Włodarczak (2015). "Temporal aspects of breathing and turn-taking in Swedish multiparty conversations". In: *Proceedings from Fonetik 2015 :* Working Papers, Department of Linguistics and Phonetics 55, pp. 47–50 (cit. on pp. 1, 21).

Han, Jing et al. (June 2021). "Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data". en. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, pp. 8328–8332. DOI: 10.1109/ICASSP39728.2021.9414576 (cit. on p. 48).

Hardcastle, William J and Alain Marchal (2012). *Speech production and speech modelling*. Vol. 55. Springer Science & Business Media (cit. on p. 1).

Harvill, John et al. (2021). "Classification of COVID-19 from cough using autoregressive predictive coding pretraining and spectral data augmentation". In: *Interspeech 2021*, pp. 926–930. DOI: 10.21437/Interspeech.2021-799 (cit. on p. 48).

Hassan, Abdelfatah, Ismail Shahin, and Mohamed Bader Alsabek (2020). "COVID-19 Detection System using Recurrent Neural Networks". In: *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, pp. 1–5. DOI: 10.1109/CCCI49893.2020.9256562 (cit. on p. 48).

Hastie, Trevor, Jerome Friedman, and Robert Tibshirani (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer. DOI: 10.1007/978-0-387-21606-5 (cit. on p. 12).

Heck, Detlef H et al. (2017). "Breathing as a fundamental rhythm of brain function". In: *Frontiers in Neural Circuits* 10. Publisher: Frontiers, p. 115 (cit. on p. 21).

Heimer, D., J. Brami, D. Lieberman, and H. Bark (June 1990). "Respiratory muscle performance in patients with type 1 diabetes". eng. In: *Diabetic Medicine: A Journal of the British Diabetic Association* 7.5, pp. 434–437. DOI: 10.1111/j.1464-5491.1990.tb01419.x (cit. on p. 95).

Henderson, Alan, Frieda Goldman-Eisler, and Andrew Skarbek (1965). "Temporal patterns of cognitive activity and breath control in speech". In: *Language and Speech* 8.4. Publisher: SAGE Publications Sage UK: London, England, pp. 236–242 (cit. on p. 21).

Hixon, Thomas J. (1987). *Respiratory function in speech and song*. College-Hill (cit. on p. 22).

Hixon, Thomas J., Jere Mead, and Michael D. Goldman (1976). "Dynamics of the chest wall during speech production: Function of the thorax, rib cage, diaphragm, and abdomen". In: *Journal of Speech and Hearing Research* 19.2. Publisher: ASHA, pp. 297–356 (cit. on p. 22).

Ho, Tin Kam (1995). "Random decision forests". In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1. IEEE, pp. 278–282 (cit. on pp. xiii, 12, 49, 52, 78).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural Computation* 9.8. Publisher: MIT Press, pp. 1735–1780 (cit. on p. 13).

Hoit, Jeannette D. and Thomas J. Hixon (Sept. 1986). "Body Type and Speech Breathing". In: *Journal of Speech, Language, and Hearing Research* 29.3. Publisher: American Speech-Language-Hearing Association, pp. 313–324. DOI: 10.1044/jshr.2903.313 (cit. on p. 21).

Hoit, Jeannette D. and Thomas J. Hixon (Sept. 1987). "Age and Speech Breathing". In: *Journal of Speech, Language, and Hearing Research* 30.3. Publisher: American Speech-Language-Hearing Association, pp. 351–366. DOI: 10.1044/jshr.3003.351 (cit. on p. 21).

Hoit, Jeannette D., Thomas J. Hixon, Mary Ellen Altman, and Wayne J. Morgan (June 1989). "Speech Breathing in Women". In: *Journal of Speech, Language, and Hearing Research* 32.2. Publisher: American Speech-Language-Hearing Association, pp. 353–365. DOI: 10.1044/jshr.3202.353 (cit. on p. 21).

Hoit, Jeannette D., Nancy Pearl Solomon, and Thomas J. Hixon (June 1993). "Effect of Lung Volume on Voice Onset Time (VOT)". In: *Journal of Speech, Language, and Hearing Research* 36.3. Publisher: American Speech-Language-Hearing Association, pp. 516–520. DOI: 10.1044/jshr.3603.516 (cit. on p. 19).

Huber, Jessica E., Bharath Chandrasekaran, and John J. Wolstencroft (2005). "Changes to respiratory mechanisms during speech as a result of different cues to increase loudness". In: *Journal of Applied Physiology* 98.6. tex.eprint:

https://doi.org/10.1152/japplphysiol.01239.2004, pp. 2177–2184. DOI: 10.1152/japplphysiol.01239.2004 (cit. on p. 19).

James, Alex Pappachen (2015). "Heart rate monitoring using human speech spectral features". In: *Human-centric Computing and Information Sciences* 5.1. Publisher: Springer, p. 33 (cit. on pp. 2, 85).

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor (2023). *An Introduction to Statistical Learning: with Applications in Python*. en. Springer Texts in Statistics. Cham: Springer International Publishing. DOI: 10.1007/978-3-031-38747-0 (cit. on p. 11).

Jati, Arindam, Paula G. Williams, Brian Baucom, and Panayiotis Georgiou (2018). "Towards Predicting Physiology from Speech During Stressful Conversations: Heart Rate and Respiratory Sinus Arrhythmia". In: *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4944–4948. DOI: 10.1109/ICASSP.2018.8461500 (cit. on pp. 1, 85).

Javanmardi, Farhad, Saska Tirronen, Manila Kodali, Sudarsana Reddy Kadiri, and Paavo Alku (June 2023). "Wav2vec-Based Detection and Severity Level Classification of Dysarthria From Speech". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10094857 (cit. on p. 11).

Jeon, Jouhyun, Adam Palanica, Sarah Sarabadani, Michael Lieberman, and Yan Fossat (Sept. 2020). *Biomarker potential of real-world voice signals to predict abnormal blood glucose levels*. en. Pages: 2020.09.25.314096 Section: New Results. DOI: 10.1101/2020.09.25.314096 (cit. on p. 96).

Kabil, Selen Hande, Hannah Muckenhirn, and Mathew Magimai.-Doss (Sept. 2018). "On learning to identify genders from raw speech signal using CNNs". In: *Interspeech 2018*, pp. 287–291. DOI: 10.21437/Interspeech.2018-1240 (cit. on pp. 13, 65).

Kabitz, H.-J. et al. (Jan. 2008). "Diabetic polyneuropathy is associated with respiratory muscle impairment in type 2 diabetes". en. In: *Diabetologia* 51.1, pp. 191–197. DOI: 10.1007/s00125-007-0856-0 (cit. on p. 95).

Khare, Aparna, Srinivas Parthasarathy, and Shiva Sundaram (2020). "Multi-modal embeddings using multi-task learning for emotion recognition". In: *Interspeech 2020*, pp. 384–388. DOI: 10.21437/Interspeech.2020-1827 (cit. on p. 63).

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A method for stochastic optimization". In: *Computing Research Repository (CoRR)* abs/1412.6980 (cit. on pp. 12, 25, 59).

Kirschbaum, Clemens, Karl-Martin Pirke, and Dirk H Hellhammer (1993). "The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting". In: *Neuropsychobiology* 28.1-2. Publisher: S. Karger AG Basel, Switzerland, pp. 76–81 (cit. on pp. 66, 89).

Klatt, D. H., K. N. Stevens, and J. Mead (1968). "Studies of articulatory activity and airflow during speech*". In: *Annals of the New York Academy of Sciences* 155.1. tex.eprint: https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.1968.tb56748.x, pp. 42–55. DOI: 10.1111/j.1749-6632.1968.tb56748.x (cit. on p. 21).

Klumpp, P. et al. (2021). “The phonetic footprint of covid-19?” In: *Interspeech 2021*, pp. 441–445. DOI: 10.21437/Interspeech.2021-1488 (cit. on pp. 48, 49).

Konno, K. and J. Mead (1967). “Measurement of the separate volume changes of rib cage and abdomen during breathing”. In: *Journal of Applied Physiology* 22.3. tex.eprint: https://doi.org/10.1152/jappl.1967.22.3.407, pp. 407–422. DOI: 10.1152/jappl.1967.22.3.407 (cit. on pp. 22, 23).

Koolagudi, Shashidhar G., Y. V. Srinivasa Murthy, and Siva P. Bhaskar (Mar. 2018). “Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition”. en. In: *International Journal of Speech Technology* 21.1, pp. 167–183. DOI: 10.1007/s10772-018-9495-8 (cit. on p. 19).

LeCun, Yann and Yoshua Bengio (1995). “Convolutional networks for images, speech, and time series”. In: *The Handbook of Brain Theory and Neural Networks* 3361.10. Publisher: Citeseer (cit. on p. 13).

Lehmann, Vera et al. (Feb. 2024). “Machine Learning to Infer a Health State Using Biomedical Signals — Detection of Hypoglycemia in People with Diabetes while Driving Real Cars”. In: *NEJM AI* 1.3. Publisher: Massachusetts Medical Society, AIoa2300013. DOI: 10.1056/AIoa2300013 (cit. on pp. xv, 98, 99).

Liu, Zheli, Zhendong Wu, Tong Li, Jin Li, and Chao Shen (July 2018). “GMM and CNN Hybrid Method for Short Utterance Speaker Recognition”. In: *IEEE Transactions on Industrial Informatics* 14.7, pp. 3244–3252. DOI: 10.1109/TII.2018.2799928 (cit. on p. 10).

Löfqvist, Anders (Oct. 1986). “The long-time-average spectrum as a tool in voice research”. In: *Journal of Phonetics*. Voice Acoustics and Dysphonia Gotland, Sweden, August 1985 14.3, pp. 471–475. DOI: 10.1016/S0095-4470(19)30692-8 (cit. on p. 10).

Loshchilov, Ilya and Frank Hutter (2019). “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations (ICLR)* (cit. on p. 68).

Maaten, Laurens van der and Geoffrey Hinton (2008). “Visualizing data using t-SNE”. In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605 (cit. on p. 61).

MacLarnon, Ann and Gwen P. Hewitt (July 1999). “The evolution of human speech: The role of enhanced breathing control”. In: *American journal of physical anthropology* 109, pp. 341–63. DOI: 10.1002/(SICI)1096-8644(199907)109:3<341::AID-AJPA5>3.0.CO;2-2 (cit. on pp. 19, 65).

Magliano, Dianna J., Edward J. Boyko, and IDF Diabetes Atlas 10th edition scientific committee (2021). *Idf diabetes atlas*. Edition: 10th ISBN: 9782930229980 Series: IDF diabetes atlas (cit. on p. 95).

Makhoul, J. (1975). “Linear prediction: A tutorial review”. In: *Proceedings of the IEEE* 63.4, pp. 561–580. DOI: 10.1109/PROC.1975.9792 (cit. on p. 23).

Makowski, Dominique et al. (Aug. 2021). “NeuroKit2: A Python toolbox for neurophysiological signal processing”. en. In: *Behavior Research Methods* 53.4, pp. 1689–1696. DOI: 10.3758/s13428-020-01516-y (cit. on p. 89).

Mamidisetti, Suresh and A. Mallikarjuna Reddy (2023). “A Stacking-based Ensemble Framework for Automatic Depression Detection using Audio Signals”. en. In: *International Jour-*

nal of Advanced Computer Science and Applications 14.7. DOI: 10.14569/IJACSA.2023.0140767 (cit. on p. 10).

Markitantov, Maxim, Denis Dresvyanskiy, Danila Mamontov, Heysem Kaya, Wolfgang Minker, and Alexey Karpov (2020). "Ensembling end-to-end deep models for computational paralinguistics tasks: ComParE 2020 mask and breathing sub-challenges". In: *Interspeech 2020*, pp. 2072–2076. DOI: 10.21437/Interspeech.2020-2666 (cit. on pp. 2, 41, 42, 49).

Mason, Llew, Jonathan Baxter, Peter Bartlett, and Marcus Frean (1999). "Boosting Algorithms as Gradient Descent". In: *Advances in Neural Information Processing Systems*. Ed. by S. Solla, T. Leen, and K. Müller. Vol. 12. MIT Press (cit. on pp. xiii, 49, 52).

Master, Suely, Noemi de Biase, Vanessa Pedrosa, and Brasília Maria Chiari (2006). "The long-term average spectrum in research and in the clinical practice of speech therapists". In: *Pró-Fono Revista de Atualização Científica* 18, pp. 111–120 (cit. on p. 10).

Matrouf, D., J.-F. Bonastre, and C. Fredouille (2006). "Effect of Speech Transformation on Impostor Acceptance". In: *ICASSP 2006 - 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 1, pp. I–I. DOI: 10.1109/ICASSP.2006.1660175 (cit. on p. 58).

McFee, Brian et al. (2015). "librosa: Audio and Music Signal Analysis in Python". en. In: Austin, Texas, pp. 18–24. DOI: 10.25080/Majora-7b98e3ed-003 (cit. on p. 75).

Mendonça, John, Francisco Teixeira, Isabel Trancoso, and Alberto Abad (2020). "Analyzing breath signals for the interspeech 2020 ComParE challenge". In: *Interspeech 2020*, pp. 2077–2081. DOI: 10.21437/Interspeech.2020-2778 (cit. on pp. 2, 41, 42).

Mesleh, Abdelwadood, Dmitriy Skopin, Sergey Baglikov, and Anas Quteishat (2012). "Heart rate extraction from vowel speech signals". In: *Journal of Computer Science and Technology* 27.6. Publisher: Springer, pp. 1243–1251 (cit. on pp. 85, 94).

Minifie, Fred et al. (1973). "Normal aspects of speech, hearing, and language." In: Publisher: ERIC (cit. on p. 1).

Miotto, Riccardo, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley (Nov. 2018). "Deep learning for healthcare: review, opportunities and challenges". en. In: *Briefings in Bioinformatics* 19.6, pp. 1236–1246. DOI: 10.1093/bib/bbx044 (cit. on p. 19).

Mitchell, Heather L., Jeannette D. Hoit, and Watson Peter J. (Feb. 1996). "Cognitive-Linguistic Demands and Speech Breathing". In: *Journal of Speech, Language, and Hearing Research* 39.1, pp. 93–104. DOI: 10.1044/jshr.3901.93 (cit. on p. 21).

Mitra, Vikramjit et al. (July 2024). *Pre-Trained Foundation Model representations to uncover Breathing patterns in Speech*. arXiv:2407.13035. DOI: 10.48550/arXiv.2407.13035 (cit. on p. 114).

Mohamed, Abdelrahman et al. (Oct. 2022). "Self-Supervised Speech Representation Learning: A Review". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6, pp. 1179–1210. DOI: 10.1109/JSTSP.2022.3207050 (cit. on p. 11).

Morise, Masanori, Fumiya Yokomori, and Kenji Ozawa (2016). "World: a vocoder-based high-quality speech synthesis system for real-time applications". In: *IEICE Transactions on Information and Systems* 99.7. Publisher: The Institute of Electronics, Information and Communication Engineers, pp. 1877–1884 (cit. on p. 58).

Mostaani, Zohreh and Mathew Magimai.-Doss (2022). "On breathing pattern information in synthetic speech". In: *Interspeech 2022*, pp. 2768–2772. DOI: 10.21437/Interspeech.2022-10271 (cit. on pp. 47, 70).

Mostaani, Zohreh, Venkata Srikanth Nallanthighal, Aki Härmä, Helmer Strik, and Mathew Magimai.-Doss (Oct. 2024). *Estimating breathing pattern from raw speech waveform and short-term speech spectrum using neural networks*. Idiap Reseach Report Idiap-RR-12-2024. Idiap (cit. on p. 21).

Mostaani, Zohreh, RaviShankar Prasad, Bogdan Vlasenko, and Mathew Magimai-Doss (2022). "Modeling of pre-trained neural network embeddings learned from raw waveform for COVID-19 infection detection". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8482–8486. DOI: 10.1109/ICASSP43922.2022.9746271 (cit. on pp. 47, 70).

Mostaani, Zohreh, Venkata Srikanth Nallanthighal, Aki Härmä, Helmer Strik, and Mathew Magimai-Doss (2021). "On The Relationship Between Speech-Based Breathing Signal Prediction Evaluation Measures and Breathing Parameters Estimation". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1345–1349. DOI: 10.1109/ICASSP39728.2021.9414756 (cit. on p. 21).

Muckenhirn, Hannah, Vinayak Abrol, Mathew Magimai-Doss, and Sébastien Marcel (2019). "Understanding and visualizing raw waveform-based CNNs". In: *Interspeech 2019*, pp. 2345–2349. DOI: 10.21437/Interspeech.2019-2341 (cit. on pp. 14, 41, 48, 74, 114).

Muckenhirn, Hannah, Pavel Korshunov, Mathew Magimai-Doss, and Sébastien Marcel (Nov. 2017). "Long-Term Spectral Statistics for Voice Presentation Attack Detection". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.11, pp. 2098–2111. DOI: 10.1109/TASLP.2017.2743340 (cit. on p. 10).

Muckenhirn, Hannah, Mathew Magimai.-Doss, and Sébastien Marcell (2018). "Towards directly modeling raw speech signal for speaker verification using CNNS". In: *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4884–4888. DOI: 10.1109/ICASSP.2018.8462165 (cit. on pp. 13, 14, 48, 65).

Muguli, Ananya et al. (2021). "DiCOVA challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics". In: *Interspeech 2021*, pp. 901–905. DOI: 10.21437/Interspeech.2021-74 (cit. on p. 48).

Nallanthighal, Venkata Srikanth, Aki Harma, Helmer Strik, and Mathew Magimai.-Doss (Aug. 2021). "Phoneme Based Respiratory Analysis of Read Speech". en. In: *2021 29th European Signal Processing Conference (EUSIPCO)*. Dublin, Ireland: IEEE, pp. 191–195. DOI: 10.23919/EUSIPCO54536.2021.9615986 (cit. on p. 82).

Nallanthighal, Venkata Srikanth, Aki Härmä, and Helmer Strik (2019). "Deep sensing of breathing signal during conversational speech". In: *Interspeech 2019*, pp. 4110–4114. DOI: 10.21437/Interspeech.2019-1796 (cit. on pp. 22, 28, 41).

Nallanthighal, Venkata Srikanth, Aki Härmä, and Helmer Strik (2020). "Speech breathing estimation using deep learning methods". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1140–1144. DOI: 10.1109/ICASSP40776.2020.9053753 (cit. on p. 41).

Nallanthighal, Venkata Srikanth, Zohreh Mostaani, Aki Härmä, Helmer Strik, and Mathew Magimai-Doss (2021). "Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings". In: *Neural Networks* 141, pp. 211–224. DOI: https://doi.org/10.1016/j.neunet.2021.03.029 (cit. on p. 21).

Niizumi, Daisuke, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino (2021). "Byol for audio: Self-supervised learning for general-purpose audio representation". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8 (cit. on p. 89).

Ntalampiras, Stavros (May 2023). "Model Ensemble for Predicting Heart and Respiration Rate From Speech". In: *IEEE Internet Computing* 27.3, pp. 15–20. DOI: 10.1109/MIC.2023.3257862 (cit. on p. 2).

Oord, Aaron van den et al. (2016). *WaveNet: A Generative Model for Raw Audio*. arXiv:1609.03499 [cs.SD]. DOI: 10.48550/arXiv.1609.03499 (cit. on pp. 56, 58).

Oppenheim, A. V. and R. W. Schafer (2004). "From frequency to quefrency: a history of the cepstrum". In: *IEEE Signal Processing Magazine* 21.5, pp. 95–106. DOI: 10.1109/MSP.2004.1328092 (cit. on p. 23).

Orlandic, Lara, Tomas Teijeiro, and David Atienza (2021). "The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms". In: *Scientific Data* 8.1. Publisher: Nature Publishing Group, pp. 1–10 (cit. on p. 48).

Orlikoff, Robert F and RJ Baken (1989). "The effect of the heartbeat on vocal fundamental frequency perturbation". In: *Journal of Speech, Language, and Hearing Research* 32.3. Publisher: ASHA, pp. 576–582 (cit. on pp. 85, 94).

Orozco-Arroyave, Juan Rafael et al. (2015). "Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases". In: *IEEE Journal of Biomedical and Health Informatics* 19.6. Publisher: IEEE, pp. 1820–1828 (cit. on p. 1).

Ou, Zhijian and Yang Zhang (2012). "Probabilistic acoustic tube: a probabilistic generative model of speech for speech analysis/synthesis". In: *Artificial Intelligence and Statistics*. PMLR, pp. 841–849 (cit. on p. 23).

Palaz, Dimitri, Ronan Collobert, and Mathew Magimai.-Doss (2013). "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks". In: *Interspeech 2013*, pp. 1766–1770 (cit. on pp. 13, 48, 65).

Palaz, Dimitri, Mathew Magimai-Doss, and Ronan Collobert (2019). "End-to-end acoustic modeling using convolutional neural networks for HMM-Based automatic speech recognition". In: *Speech Communication* 108, pp. 15–32. DOI: 10.1016/j.specom.2019.01.004 (cit. on pp. 14, 41).

Paszke, Adam et al. (2019). "PyTorch: An imperative style, high-performance deep learning library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 8024–8035 (cit. on pp. 24, 59).

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in python". In: *Journal of Machine Learning Research* 12.Oct, pp. 2825–2830 (cit. on pp. 52, 78, 102).

Pieniawska, A., A. Horodnicka-Józwa, E. Petriczko, and M. Walczak (2012). “Evaluation of respiratory function tests in children and adolescents with type 1 diabetes”. Polish. In: *Pediatric Endocrinology, Diabetes and Metabolism* 18.1, pp. 15–20 (cit. on p. 95).

Pokorny, Florian B., Franz Graf, Franz Pernkopf, and Björn W Schuller (Sept. 2015). “Detection of negative emotions in speech signals using bags-of-audio-words”. In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. ISSN: 2156-8111, pp. 879–884. DOI: 10.1109/ACII.2015.7344678 (cit. on p. 8).

Pompe, Simone, Adria Mallol-Ragolta, Nicolas Schauer, and Björn W Schuller (2023). “Exploring Shapely Values for Blood Glucose Level Prediction from Speech”. In: *Speech Communication; 15th ITG Conference*, pp. 81–85. DOI: 10.30420/456164015 (cit. on pp. 96, 97).

Purohit, Tilak, Imen Ben Mahmoud, Bogdan Vlasenko, and Mathew Magimai.-Doss (2022). “Comparing supervised and self-supervised embedding for ExVo Multi-Task learning track”. In: *Proceedings of the ICML Expressive Vocalizations (ExVo) Workshop and Competition 2022* (cit. on p. 69).

Purohit, Tilak, Bogdan Vlasenko, and Mathew Magimai.-Doss (Aug. 2023). “Implicit phonetic information modeling for speech emotion recognition”. en. In: *Interspeech 2023*, pp. 1883–1887. DOI: 10.21437/Interspeech.2023-1999 (cit. on p. 11).

Purohit, Tilak, Sarthak Yadav, Bogdan Vlasenko, S. Pavankumar Dubagunta, and Mathew Magimai.-Doss (June 2023). “Towards Learning Emotion Information from Short Segments of Speech”. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095892 (cit. on pp. 13, 67).

Qi, Jun, Jun Du, Sabato Marco Siniscalchi, and Chin-Hui Lee (2019). “A theory on deep neural network based vector-to-vector regression with an illustration of its expressive power in speech enhancement”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12. Publisher: IEEE, pp. 1932–1943 (cit. on p. 23).

Qi, Jun, Jun Du, Sabato Marco Siniscalchi, Xiaoli Ma, and Chin-Hui Lee (2020). “Analyzing upper bounds on mean absolute errors for deep neural network based vector-to-vector regression”. In: *IEEE Transactions on Signal Processing*. Publisher: IEEE (cit. on pp. 23, 39).

Ravanelli, Mirco and Yoshua Bengio (2018). “Speaker recognition from raw waveform with sincnet”. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 1021–1028 (cit. on p. 65).

Rethage, D., J. Pons, and X. Serra (2018). “A wavenet for speech denoising”. In: *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5069–5073. DOI: 10.1109/ICASSP.2018.8462417 (cit. on p. 23).

Ritwik, Kotra Venkata Sai, Shareef Babu Kalluri, and Deepu Vijayasenan (2021). “COVID-19 detection from spectral features on the DiCOVA dataset”. In: *Interspeech 2021*, pp. 936–940. DOI: 10.21437/Interspeech.2021-1031 (cit. on p. 48).

Ruinskiy, D. and Y. Lavner (Mar. 2007). “An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals”. In: *IEEE Transactions on*

Audio, Speech, and Language Processing 15.3, pp. 838–850. DOI: 10.1109/TASL.2006.889750 (cit. on p. 20).

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (Oct. 1986). "Learning representations by back-propagating errors". en. In: *Nature* 323.6088. Publisher: Nature Publishing Group, pp. 533–536. DOI: 10.1038/323533a0 (cit. on pp. 12, 13).

Russell, James A (1980). "A circumplex model of affect." In: *Journal of Personality and Social Psychology* 39.6. Publisher: American Psychological Association, p. 1161 (cit. on p. 63).

Ryskaliyev, Aibek, Sanzhar Askaruly, and Alex Pappachen James (2016). "Speech signal analysis for the estimation of heart rates under different emotional states". In: *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, pp. 1160–1165 (cit. on p. 85).

Saghiri, Mohammad Ali, Anna Vakhnovetsky, and Julia Vakhnovetsky (Mar. 2022). "Scoping review of the relationship between diabetes and voice quality". In: *Diabetes Research and Clinical Practice* 185, p. 109782. DOI: 10.1016/j.diabres.2022.109782 (cit. on p. 95).

Sainath, Tara N., Ron J. Weiss, Andrew Senior, Kevin W. Wilson, and Oriol Vinyals (2015). "Learning the speech front-end with raw waveform CLDNNs". In: *Interspeech 2015*, pp. 1–5. DOI: 10.21437/Interspeech.2015-1 (cit. on p. 65).

Satt, Aharon, Shai Rozenberg, and Ron Hoory (Aug. 2017). "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms". en. In: *Interspeech 2017*, pp. 1089–1093. DOI: 10.21437/Interspeech.2017-200 (cit. on p. 10).

Scheidwasser-Clow, Neil, Mikolaj Kegler, Pierre Beckmann, and Milos Cernak (2022). "SERAB: A multi-lingual benchmark for speech emotion recognition". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing ((ICASSP)*. IEEE, pp. 7697–7701 (cit. on p. 86).

Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview". In: *Neural Networks* 61, pp. 85–117. DOI: https://doi.org/10.1016/j.neunet.2014.09.003 (cit. on p. 25).

Schmitt, Maximilian and Björn W Schuller (2017). "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit". In: *Journal of Machine Learning Research* 18.96, pp. 1–5 (cit. on pp. 52, 59).

Scholkmann, Felix, Jens Boss, and Martin Wolf (2012). "An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals". In: *Algorithms* 5.4. Publisher: Multidisciplinary Digital Publishing Institute, pp. 588–603 (cit. on pp. 22, 23).

Schuller, Björn W (2018). "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends". In: *Communications of the ACM* 61.5. Publisher: ACM New York, NY, USA, pp. 90–99 (cit. on p. 63).

Schuller, Björn W, Felix Friedmann, and Florian Eyben (2013). "Automatic recognition of physiological parameters in the human voice: Heart rate and skin conductance". In: *ICASSP 2013 - 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7219–7223 (cit. on p. 85).

Schuller, Björn W, Felix Friedmann, and Florian Eyben (2014). "The munich biovoice corpus: Effects of physical exercising, heart rate, and skin conductance on human speech pro-

duction." In: *Proceedings 9th Language Resources and Evaluation Conference (LREC 2014)*, pp. 1506–1510 (cit. on pp. 85, 94).

Schuller, Björn W, Stefan Steidl, Anton Batliner, et al. (2013). "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism". In: *Interspeech 2013*, pp. 148–152. DOI: 10.21437/Interspeech.2013-56 (cit. on pp. 9, 89).

Schuller, Björn W, Bogdan Vlasenko, Dejan Arsic, Gerhard Rigoll, and Andreas Wendemuth (2008). "Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition". In: *2008 IEEE International Conference on Multimedia and Expo*, pp. 1333–1336. DOI: 10.1109/ICME.2008.4607689 (cit. on p. 65).

Schuller, Björn W et al. (Sept. 2015). "The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, Parkinson's & eating condition". en. In: *Interspeech 2015*, pp. 478–482. DOI: 10.21437/Interspeech.2015-179 (cit. on p. 9).

Schuller, Björn W et al. (2020). "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks". en. In: *Interspeech 2020*, pp. 2042–2046. DOI: 10.21437/Interspeech.2020-32 (cit. on pp. 2, 22, 28, 29, 41–43).

Schuller, Björn W et al. (2021). "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates". In: *Interspeech 2021*, pp. 431–435. DOI: 10.21437/Interspeech.2021-19 (cit. on pp. 9, 48, 49, 51).

Sebastian, Jilt, Mriganka Sur, Hema A. Murthy, and Mathew Magimai.-Doss (2020). "Signal-to-signal neural networks for improved spike estimation from calcium imaging data". In: *bioRxiv : the Preprint Server for Biology*. Publisher: Cold Spring Harbor Laboratory tex.elocation-id: 2020.05.01.071993 tex.eprint: https://www.biorxiv.org/content/early/2020/10/18/2020.05.01.071993.full.pdf. DOI: 10.1101/2020.05.01.071993 (cit. on p. 23).

Second DiCOVA challenge (2021). URL: https://dicovachallenge.github.io/ (visited on 2021) (cit. on pp. 49, 50, 52).

Sejdić, Ervin, Igor Djurović, and Jin Jiang (2009). "Time–frequency feature representation using energy concentration: An overview of recent advances". In: *Digital Signal Processing* 19.1, pp. 153–183 (cit. on p. 24).

Shah, Mohit, Ming Tu, Visar Berisha, Chaitali Chakrabarti, and Andreas Spanias (2019). "Articulation constrained learning with application to speech emotion recognition". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2019.1. Publisher: Springer, pp. 1–17 (cit. on p. 65).

Sharma, Neeraj et al. (2020). "Coswara — A database of breathing, cough, and voice sounds for COVID-19 diagnosis". In: *Interspeech 2020*, pp. 4811–4815. DOI: 10.21437/Interspeech.2020-2768 (cit. on pp. 48, 50).

Shen, Guang et al. (2020). "WISE: Word-level interaction-based multimodal fusion for speech emotion recognition". In: *Interspeech 2020*, pp. 369–373. DOI: 10.21437/Interspeech.2020-3131 (cit. on p. 63).

Shor, Joel et al. (2020). "Towards learning a universal non-semantic representation of speech". In: *Interspeech 2020*, pp. 140–144. DOI: 10.21437/Interspeech.2020-1242 (cit. on p. 86).

Sidorova, Julia, Pablo Carbonell, and Milena Čukić (Sept. 2022). "Blood Glucose Estimation From Voice: First Review of Successes and Challenges". In: *Journal of Voice* 36.5, 737.e1–737.e10. DOI: 10.1016/j.jvoice.2020.08.034 (cit. on p. 95).

Slifka, Janet (2006). "Some physiological correlates to regular and irregular phonation at the end of an utterance". In: *Journal of Voice* 20.2, pp. 171–186. DOI: https://doi.org/10.1016/j.jvoice.2005.04.002 (cit. on p. 19).

Smith, Jennifer, Andreas Tsiartas, Elizabeth Shriberg, Andreas Kathol, Adrian Willoughby, and Massimiliano De Zambotti (2017). "Analysis and prediction of heart rate using speech features from natural speech". In: *ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 989–993 (cit. on pp. 1, 2, 85).

Smith, Lindsey K. and Alexander M. Goberman (2014). "Long-time average spectrum in individuals with Parkinson disease". eng. In: *NeuroRehabilitation* 35.1, pp. 77–88. DOI: 10.3233/NRE-141102 (cit. on p. 10).

Södergren, Isabella, Maryam Pahlavan Nodeh, Prakash Chandra Chhipa, Konstantina Nikolaidou, and György Kovács (2021). "Detecting COVID-19 from audio recording of coughs using random forests and support vector machines". In: *Interspeech 2021*, pp. 916–920. DOI: 10.21437/Interspeech.2021-2191 (cit. on p. 49).

Solomon, Nancy Pearl and Thomas J Hixon (1993). "Speech breathing in Parkinson's disease". In: *Journal of Speech, Language, and Hearing Research* 36.2. Publisher: ASHA, pp. 294–310 (cit. on p. 21).

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *The Journal of Machine Learning Research* 15.1. Publisher: JMLR. org, pp. 1929–1958 (cit. on p. 68).

Stappen, Lukas, Alice Baird, et al. (Oct. 2021). "The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress". In: *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. MuSe '21. New York, NY, USA: Association for Computing Machinery, pp. 5–14. DOI: 10.1145/3475957.3484450 (cit. on pp. 2, 89).

Stappen, Lukas, Lea Schumann, et al. (2021). "MuSe-Toolbox: The Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox". In: *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. MuSe '21. Virtual Event, China: Association for Computing Machinery, pp. 75–82. DOI: 10.1145/3475957.3484451 (cit. on p. 63).

Stappen, Lukas et al. (Oct. 2020). "MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media: Emotional Car Reviews in-the-wild". In: *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*. MuSe'20. New York, NY, USA: Association for Computing Machinery, pp. 35–44. DOI: 10.1145/3423327.3423673 (cit. on p. 63).

Stevens, S. S., J. Volkmann, and E. B. Newman (1937). "A scale for the measurement of the psychological magnitude pitch". In: *The Journal of the Acoustical Society of America* 8.3.

tex.eprint: https://doi.org/10.1121/1.1915893, pp. 185–190. DOI: 10.1121/1.1915893 (cit. on p. 24).

Székely, Éva, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson (2020). "Breathing and Speech Planning in Spontaneous Speech Synthesis". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7649–7653. DOI: 10.1109/ICASSP40776.2020.9054107 (cit. on p. 21).

Tanaka, Kou, Takuhiro Kaneko, Nobukatsu Hojo, and Hirokazu Kameoka (2018). "Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks". In: *IEEE Spoken Language Technology Workshop (SLT)*, pp. 632–639. DOI: 10.1109/SLT.2018.8639636 (cit. on p. 58).

Taylor, Paul (2009). *Text-to-speech synthesis*. tex.optdoi: 10.1017/CBO9780511816338. Cambridge: Cambridge University Press (cit. on p. 56).

Teixeira, João Paulo, Carla Oliveira, and Carla Lopes (2013). "Vocal acoustic analysis–jitter, shimmer and hnr parameters". In: *Procedia Technology* 9. Publisher: Elsevier, pp. 1112–1122 (cit. on p. 19).

Todisco, Massimiliano et al. (2019). "ASVspoof 2019: Future horizons in spoofed and fake audio detection". In: *Interspeech 2019*, pp. 1008–1012. DOI: 10.21437/Interspeech.2019-2249 (cit. on pp. 57, 58, 62).

Trigeorgis, George et al. (Mar. 2016). "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network". en. In: *ICASSP 2016 - 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE, pp. 5200–5204. DOI: 10.1109/ICASSP.2016.7472669 (cit. on p. 48).

Tschope, C., F. Duckhorn, M. Wolff, and G. Saeltzer (Dec. 2015). "Estimating Blood Sugar from Voice Samples: A Preliminary Study". en. In: *2015 International Conference on Computational Science and Computational Intelligence (CSCI)*. Las Vegas, NV, USA: IEEE, pp. 804–805. DOI: 10.1109/CSCI.2015.184 (cit. on p. 96).

Turian, Joseph et al. (Dec. 2022). "HEAR: Holistic Evaluation of Audio Representations". In: *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*. Ed. by Douwe Kiela, Marco Ciccone, and Barbara Caputo. Vol. 176. Proceedings of Machine Learning Research. PMLR, pp. 125–145 (cit. on p. 86).

Tzirakis, Panagiotis, Anh Nguyen, Stefanos Zafeiriou, and Björn W Schuller (2021). "Speech emotion recognition using semantic information". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6279–6283 (cit. on p. 63).

Tzirakis, Panagiotis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou (2017). "End-to-end multimodal emotion recognition using deep neural networks". In: *IEEE Journal of Selected Topics in Signal Processing* 11.8. Publisher: IEEE, pp. 1301–1309 (cit. on p. 63).

Ulanovsky, Yakov Benediktovich, Aleksandr Mihaylovich Frolov, Alena Yakovlevna Kozlova, and Maksim Aleksandrovich Fatkin (May 2014). "Device for blood glucose level determination". en. Patent WO2014072823A2 (cit. on pp. 1, 95).

Usman, Mohammed et al. (2021). "Heart rate detection and classification from speech spectral features using machine learning". In: *Archives of Acoustics* 46.1, pp. 41–53. DOI: 10.24425/aoa.2021.136559 (cit. on p. 85).

Vapnik, Vladimir (1963). "Pattern recognition using generalized portrait method". In: *Automation and Remote Control* 24.6, pp. 774–780 (cit. on p. 12).

Vásquez-Correa, J. C., Julian Fritsch, Juan Rafael Orozco-Arroyave, Elmar Nöth, and Mathew Magimai-Doss (Aug. 2021). "On Modeling Glottal Source Information for Phonation Assessment in Parkinson's Disease". en. In: *Interspeech 2021*, pp. 26–30. DOI: 10.21437/Interspeech.2021-1084 (cit. on p. 3).

Vásquez-Correa, J. C., Juan Rafael Orozco-Arroyave, and Elmar Nöth (Aug. 2017). "Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease". en. In: *Interspeech 2017*, pp. 314–318. DOI: 10.21437/Interspeech.2017-1078 (cit. on p. 10).

Vergin, Rivarol, Azarshid Farhat, and Douglas O'Shaughnessy (1996). "Robust gender-dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification". In: *Proceeding of 4th International Conference on Spoken Language Processing. ICSLP'96*. Vol. 2. IEEE, pp. 1081–1084 (cit. on pp. 1, 65).

Virtanen, Pauli et al. (Mar. 2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python". en. In: *Nature Methods* 17.3. Publisher: Nature Publishing Group, pp. 261–272. DOI: 10.1038/s41592-019-0686-2 (cit. on p. 76).

Vlasenko, Bogdan, David Philippou-Hübner, Dmytro Prylipko, Ronald Böck, Ingo Siegert, and Andreas Wendemuth (2011). "Vowels formants analysis allows straightforward detection of high arousal emotions". In: *2011 IEEE International Conference on Multimedia and Expo*. IEEE, pp. 1–6 (cit. on p. 65).

Von Euler, C (1982). "Some aspects of speech breathing physiology". In: *Speech Motor Control*. Elsevier, pp. 95–103 (cit. on p. 19).

Wang, Yu-Tsai, Jordan R Green, Ignatius SB Nip, Ray D Kent, and Jane Finley Kent (2010). "Breath group analysis for reading and spontaneous speech in healthy adults". In: *Folia Phoniatrica et Logopaedica* 62.6. Publisher: Karger Publishers, pp. 297–302 (cit. on p. 21).

Wang, Yuxuan et al. (2017). "Tacotron: Towards end-to-end speech synthesis". In: *Interspeech 2017*, pp. 4006–4010. DOI: 10.21437/Interspeech.2017-1452 (cit. on p. 56).

WHO Coronavirus (2021). URL: https://www.who.int/health-topics/coronavirus (visited on 2021) (cit. on p. 48).

Williams, Carl E and Kenneth N Stevens (1972). "Emotions and speech: Some acoustical correlates". In: *The Journal of the Acoustical Society of America* 52.4B. Publisher: Acoustical Society of America, pp. 1238–1250 (cit. on p. 85).

Winkworth, Alison L, Pamela J Davis, Elizabeth Ellis, and Roger D Adams (1994). "Variability and consistency in speech breathing during reading: Lung volumes, speech intensity, and linguistic factors". In: *Journal of Speech, Language, and Hearing Research* 37.3. Publisher: ASHA, pp. 535–556 (cit. on p. 21).

Włodarczak, Marcin and Mattias Heldner (May 2017). "Respiratory Constraints in Verbal and Non-verbal Communication". In: *Frontiers in Psychology* 8. tex.pmcid: PMC5434352. DOI: 10.3389/fpsyg.2017.00708 (cit. on pp. 1, 19, 22).

Włodarczak, Marcin, Mattias Heldner, and Jens Edlund (2015). "Breathing in conversation : An unwritten history". In: *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication :* Linköping electronic conference proceedings. Number: 110. Stockholm University, Phonetics, pp. 107–112 (cit. on p. 19).

Wu, Haiping et al. (Oct. 2021). "CvT: Introducing Convolutions to Vision Transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22–31 (cit. on p. 90).

Wu, Wen, Chao Zhang, and Philip C. Woodland (June 2023). "Self-Supervised Representations in Speech-Based Depression Detection". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10094910 (cit. on p. 11).

Wu, Zhizheng, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li (2015). "Spoofing and countermeasures for speaker verification: A survey". In: *Speech Communication* 66, pp. 130–153. DOI: https://doi.org/10.1016/j.specom.2014.10.005 (cit. on p. 56).

Xu, Yong, Jun Du, Li-Rong Dai, and Chin-Hui Lee (2014). "A regression approach to speech enhancement based on deep neural networks". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.1. Publisher: IEEE, pp. 7–19 (cit. on p. 23).

Yadav, Sarthak, Tilak Purohit, Zohreh Mostaani, Bogdan Vlasenko, and Mathew Magimai.-Doss (Oct. 2022). "Comparing Biosignal and Acoustic feature Representation for Continuous Emotion Recognition". In: *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. MuSe' 22. New York, NY, USA: Association for Computing Machinery, pp. 37–45. DOI: 10.1145/3551876.3554812 (cit. on pp. 47, 64, 68).

Yang, Shu-wen et al. (2021). *SUPERB: Speech processing Universal PERformance Benchmark*. arXiv:2105.01051 [cs.CL]. DOI: 10.48550/arXiv.2105.01051 (cit. on p. 86).

Yuan, Jiahong, Xingyu Cai, Renjie Zheng, Liang Huang, and Kenneth Church (2021). *The Role of Phonetic Units in Speech Emotion Recognition*. arXiv:2108.01132 [cs.CL]. DOI: 10.48550/arXiv.2108.01132 (cit. on p. 65).

Zeghidour, Neil, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi (2021). "{LEAF}: a learnable frontend for audio classification". In: *International Conference on Learning Representations* (cit. on p. 65).

Zen, Heiga, Keiichi Tokuda, and Alan W. Black (2009). "Statistical parametric speech synthesis". In: *Speech Communication* 51.11. tex.optdoi: 10.1016/j.specom.2009.04.004, pp. 1039–1064 (cit. on p. 56).

Zhang, Jianhua, Zhong Yin, Peng Chen, and Stefano Nichele (2020). "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review". In: *Information Fusion* 59. Publisher: Elsevier, pp. 103–126 (cit. on p. 63).

Zhang, Qiang, Xianxiang Chen, Qingyuan Zhan, Ting Yang, and Shanhong Xia (Nov. 2017). "Respiration-based emotion recognition with deep learning". en. In: *Computers in Industry* 92-93, pp. 84–90. DOI: 10.1016/j.compind.2017.04.005 (cit. on p. 65).

Zineldin, Mokhles Abdel Fadil, Kamel Abdel Ghaffar Hasan, and Ahmed Salama Al-Adl (Jan. 2015). "Respiratory function in type II diabetes mellitus". In: *Egyptian Journal of Chest Diseases and Tuberculosis* 64.1, pp. 219–223. DOI: 10.1016/j.ejcdt.2014.08.008 (cit. on p. 95).

Zwald, Laurent and Sophie Lambert-Lacroix (2012). *The BerHu penalty and the grouped effect*. arXiv:1207.6868 [math.ST]. DOI: 10.48550/arXiv.1207.6868 (cit. on p. 27).

Zohreh Mostaani

Martigny, Switzerland
+41779520084
mostaani.zohreh@gmail.com
zohreh-mostaani
t2VMpU8AAAAJ

Experience

Feb 2020 – Dec 2024 **Research Assistant**, *Idiap Research Institute*, Martigny, Switzerland

Integrated physiological signals in speech processing technologies using machine learning and deep learning.

- Pioneered developing deep learning models to estimate breathing signals from raw speech waveforms using TensorFlow. Achieved a 25% reduction in input lag compared to spectral-based models.
- Developed models for COVID-19 detection, presentation attack detection, emotion recognition, heart rate estimation, and hypoglycemia detection models in PyTorch and Scikit-learn using data-driven and hand-crafted features extracted from speech signals.
- Compiled two databases with simultaneous recordings of speech and physiological signals from more than 30 participants.
- Guided two Master's students through internships and thesis projects. Supported Master's-level courses *Machine Learning* and *Introduction to Speech Processing* as a teaching assistant, providing help with assignments, projects, coding challenges, and grading.

Apr 2019 – Jan 2020 **Research and Development Engineer**, *Idiap Research Institute*, Martigny, Switzerland

- Documented, tested, and added new functionalities to BEAT framework; a framework enabling the definition, execution, and comparison of data-driven workflows: Python, Django, Sphinx.
- Devised a protocol for collecting face biometric data in automotive environments.

May 2017 – Mar 2019 **Internship**, *Idiap Research Institute*, Martigny, Switzerland

- Collected two extensive face biometric datasets from over 100 participants, including protocol design for data acquisition and generation of face presentation attacks.

Dec 2016 – May 2017 **Internship**, *Idiap Research Institute*, Martigny, Switzerland

Incorporated an Automatic Speech Recognition (*ASR*) system into a web-based framework using Docker and RESTful API.

Jun 2016 – Nov 2016 **Internship**, *Idiap Research Institute*, Martigny, Switzerland

Implemented a Convolutional Neural Network for *Gaze Estimation* in Theano using data collected at Idiap.

Feb 2012 – Jan 2015 **Research Assistant**, *Özyegin University*, Istanbul, Turkey

- Enhanced system efficiency in free-space optical communication by achieving a 50% improvement in signal-to-noise ratio (SNR) through the use of multiple relays.
- Managed assignments, labs, and exams for Bachelor's and Master's level courses, including Physics, Mathematics, Linear Algebra, and Wireless Communications.

Education

Dec 2024 **PhD student**, *Electrical and Electronics Engineering*, **EPFL University**, Switzerland

Jan 2015 **Master of Science**, *Electrical and Electronics Engineering*, **Özyegin University**, Turkey

Sep 2011 **Bachelor of Science**, *Electrical Engineering*, **University of Tehran**, Iran

Selected Course Work

Feb 2021 – Jul 2021 **Fundamentals of Statistical Pattern recognition**, *EPFL*

Completed projects on machine learning concepts such as linear regression, logistic regression, neural networks, PCA, LDA, K-Means, GMMs, and SVMs.

Sep 2020 – Jan 2021 **Speech and Audio Coding**, *EPFL*

Studied sound production and perception mechanisms, sound signal processing such as sampling, quantization, spectral analysis, and speech coding such as linear predictive coding (LPC) and MPEG-1 layer 3 (mp3).

Feb 2020 – Jul 2020 **Statistical Sequence Processing**, *EPFL*

Gained knowledge in statistical pattern recognition using supervised and unsupervised learning, statistical sequence modeling using Markov Models.

Skills

Programming	Python, MATLAB, Shell scripting.
Toolboxes	PyTorch, Tensorflow, Scikit-learn, Pandas, Numpy, Matplotlib.
Computing	GNU/Linux, Git, Docker, Conda.
Languages	Persian (native), English (fluent), French (intermediate).

Academic Experience

Academic Service

Reviewer — IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Publications

Published over 15 papers in top-tier journals and conferences such as: **Neural networks**, **ICASSP**, and **INTERSPEECH** with more than 500 citations and an h-index of 8 (Google Scholar).

Honors and Awards

Jun 2007	Ranked 206^{th} among approximately 400,000 participants in the nationwide university entrance exam, Iran.
Mar 2006	Won acceptance in first stage of national Mathematics Olympiad and national Literature Olympiad, Iran.
Mar 2005	Won acceptance in first stage of national Mathematics Olympiad, Iran.