EPFL

Thèse n° 11387

Advancing Neural Representations for Paralinguistic Analysis: From Speech Emotion to Parkinson's Disease Assessment

Présentée le 13 février 2026

Faculté des sciences et techniques de l'ingénieur
Laboratoire de traitement des signaux 5
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Tilak PUROHIT

Acceptée sur proposition du jury

Dr J.-M. Odobez, président du jury
Prof. J.-Ph. Thiran, Dr M. Magimai Doss, directeurs de thèse
Prof. I. Trancoso, rapporteuse
Prof. S. Narayanan, rapporteur
Prof. M. Shoaran, rapporteuse

École polytechnique fédérale de Lausanne

2026

The most exciting phrase to hear in science, the one that heralds new discoveries, is not ‘Eureka!’ but ‘That’s funny..’

— Isaac Asimov

With sincere gratitude to everyone whose kindness and motivation sustained this work.

Acknowledgements

What began as an inspiring keynote by Prof. Hervé Bourlard at Interspeech 2018 has culminated in the completion of my PhD at the laboratory he formerly directed; it has been an incredible journey ever since.

I would like to begin by expressing my deepest gratitude to my supervisor, Dr. Mathew Magimai Doss, for providing me with the opportunity to pursue my PhD at Idiap. I am sincerely thankful for his guidance, mentorship, and unwavering support throughout my doctoral studies. I have learned a great deal from him and have always appreciated his enthusiasm for the field. Time and again, he reminded me of the broader scientific questions and ensured that I remained on the right track. During our meetings, Mathew would consistently emphasize, *"Tilak, what is the question you are really asking?"*—a remark that I will carry with me throughout my career. I am also deeply grateful to my thesis director, Prof. Jean-Philippe Thiran, for his constant support and encouragement.

I am sincerely thankful to the members of my thesis jury, Prof. Isabel Trancoso, Prof. Shrikanth Narayanan, and Prof. Mahsa Shoran, for agreeing to serve as experts and for the time and effort they devoted to evaluating this work. I would also like to thank Dr. Jean-Marc Odobez for serving as the jury president. The insightful discussions during the oral exam were both a privilege and a source of inspiration for the future research directions.

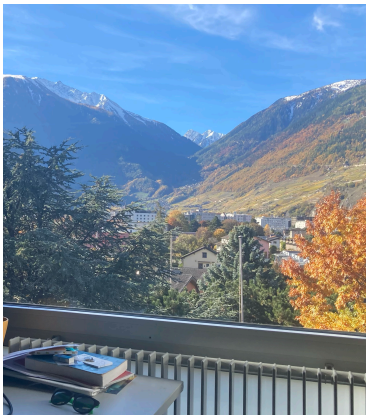
I would also like to acknowledge my mentors in India, who laid the foundation of my doctoral journey. Prof. V. Ramasubramanian, my master's supervisor at IIIT Bangalore, and Prof. Prasanta Kumar Ghosh, my supervisor at the SPIRE Lab, IISc Bangalore, introduced me to both the rigor and the joy of true research. I am grateful for their mentorship and support. I also deeply appreciate the mentorship of Dr. Ankush Mittal during my undergraduate studies at Graphic Era University in Dehradun.

I am thankful to Prof. J.R. Orozco-Arroyave (Rafa) and the GITA Lab team at UdeA, for a productive three-month stay in Medellín, Colombia that led to the work in Chapter 6. I will always cherish the culture and warmth I experienced there (¡Qué chimba es Colombia!). Additionally, I am grateful to Dr. Michael Owen and Dr. Andrew Fletcher at Amazon in Boston. My internship offered a first look into industry, teaching me how to navigate the complexities of real-world solutions and cross-team collaboration in a fast-paced environment.

I owe a debt of gratitude to my collaborators: Pavan, Sarthak, Felipe, Zohreh, Imen, and Parvaneh for their invaluable contributions to our shared publications. I also wish to highlight

the contribution of Barbara, whose Master's thesis I had the pleasure of co-supervising and which forms the foundation of Chapter 7 of this thesis.
A special and a huge shout-out to my *semi-supervisor* Dr. Bogdan Vlasenko, beyond sharing his enthusiasm and knowledge of emOtions, his friendship has been one of the highlights of my time at Idiap.
My four years in Martigny have been, and will always remain, an unforgettable chapter of my life. I have always felt blessed to pursue a PhD at Idiap, where the peacefulness of the Swiss Alps meet the brilliance of a vibrant research community. This journey was as much about scientific growth as it was about personal growth and cultural exchange (shoutout to CCDS!). I owe so much to the friends who helped me grow as a person. Rather than listing them by name, I want to honor the *fellas* through the shared experiences and memories that truly defined this journey. I will forever cherish the activities we shared- from the discipline of our early morning swims, long-distance runs, and the grit of our half-marathons, to the sheer joy of carving down the pistes, hitting the pump-tracks, and even those few strenuous but fun bouldering sessions. I will never forget the summits we conquered together-Catogne, Grand Chavalard, and Pierre Avoi, to name a few of my favorites. I will also miss our long coffee breaks, beautiful hikes for stargazing, the chaotic magic AI aka Montreux Jazz, summer volleyball sessions, and the warmth of Christmas-market vin chaud. From many conference trips across the globe to spontaneous travels around Europe, these shared moments have defined this chapter of my life.
To all past and present members of the Idiap community during my time here, the Room 303 squad, the SAS group, the *crazy* pre-Aubert-ians, the Sunset/Barrock/Apothi/Gouttière crew, the one-time-only Balélec gang and the festive cooking group: I carry such fond memories of you all. Thank you for your unlimited support and for making these four years so enriching and memorable.
Finally, I would like to thank my parents and my sister for their unconditional love and support. Even from a distance, they stood by me through all the ups and downs of this journey, and their encouragement was my constant strength.
This research was made possible through the generous support of Innosuisse and the Swiss National Science Foundation (SNSF) through the Bridge Discovery project EMIL: Emotion in the loop - a step towards a comprehensive closed-loop deep brain stimulation in Parkinson's disease (grant no. $40B2-0\_194794$). I gratefully acknowledge the financial and institutional support that enabled me to carry out this work and to participate in conferences, collaborations, and research visits that were essential to my doctoral training.

Martigny, January 16th 2026 Tilak Purohit



From my desk at Idiap (in room 303), countless hours of research and reflection were spent with this view.

Abstract

Paralinguistics derives from the Greek preposition $\pi\alpha\rho\acute{\alpha}$ (para), meaning “alongside linguistics". Paralinguistics is the field that quantifies the rich, extra-linguistic information in speech, such as affect, gender, or health. In essence, it is more concerned with *how* something is said, than *what* is said. Paralinguistic analysis examines human speech through two temporal perspectives: *states*, which represent short-term affective variations (such as emotions), and *traits*, which characterize long-term attributes, including pathological conditions (e.g. Parkinson's Disease). Stable traits acoustically shape and constrain the expression and perception of transient states; this forms an interesting continuum in paralinguistics.

Conventional Speech Emotion Recognition (SER) approaches typically rely on: (a) suprasegmental modeling of handcrafted acoustic descriptors, and (b) modeling long-duration speech signal (typically 4-6 seconds) directly using deep neural networks (DNNs). In contrast, we introduce a short-segment raw-waveform modelling strategy, demonstrating that emotion-discriminative information can be effectively captured from ≈250 millisecond(ms) speech segments using an end-to-end Convolutional Neural Network (CNN). Evaluated over various emotion corpora, the model, trained directly on raw waveforms, performed comparably to recent and conventional utterance-based modeling systems and outperformed handcrafted features extracted over the same 250ms duration. Interpretability analysis of our CNN, conducted using relevance signal analysis, revealed that the model captures emotion-relevant cepstral features, demonstrating the advantage of learning task-specific emotion cues directly from data.

Building on the finding that emotion-discriminative cues can be modeled from short speech segments that may encapsulate phoneme or grapheme level information we proposed a novel phonetically aware neural modeling framework. This framework leverages neural-representations from networks trained on subword classification to extract emotion-related cues, enabling an investigation into the usefulness of phonetically aware neural representations. Across benchmark corpora, this approach consistently outperformed conventional handcrafted acoustic features.

Drawing on the finding that data-driven neural representations can effectively capture paralinguistic states such as emotions, we extended these advances to model persistent neurological traits, focusing on Parkinson's disease (PD). Within the framework of Speech Foundation

Models (SFMs) pretrained on large amounts of healthy speech data, we proposed and validated parameter-efficient adaptation strategies for low-resource PD speech detection. We observed that the layer selection method matched the performance of full fine-tuning while requiring significantly fewer parameters. Notably, the application of Low-Rank Adaptation (LoRA) to the Whisper model surpassed other methods, demonstrating that models pretrained for task specific speech recognition are highly conducive to efficient adaptation for PD speech detection.

Finally, to examine the interaction between traits and states, we addressed the challenging task of detecting comorbid depression in PD a state whose expression is overlapping PD speech symptoms. In this complex, low data-resource setting, large-scale SFM approaches failed, exhibiting high bias and poor sensitivity. Instead, a more conventional, interpretable approach using handcrafted features and a robust feature selection method proved significantly more effective and superior. Our analysis using handcrafted features revealed that the acoustic descriptors of comorbid depression in PD differ from those of typical depression. Specifically, classifiers tend to focus on source-related features in non-PD speech, whereas both source and system features play a role in PD speech.

Collectively, this thesis introduces novel methods for fine-grained emotion recognition and parameter-efficient strategies for PD speech analysis. It offers a critical and nuanced perspective, emphasizing that although deep learning models exhibit strong representational power, their application to complex, comorbid, and low-resource scenarios demands a pragmatic, task-aware approach. While handcrafted acoustic features enable interpretability and allow detailed analysis, such transparency is not yet readily attainable with neural-representations (derived from SFMs). Overall, this work advances the understanding of how paralinguistic information is encoded in neural representations, bridging interpretability and scalability toward the development of robust, explainable models for paralinguistic *States* and *Traits* inference.

Keywords: paralinguistics, neural representations, speech emotion recognition, Parkinson's Disease, comorbid depression, speech foundation models, transfer learning, low-resource learning, parameter-efficient adaptation, machine learning

Contents

List of Figures

List of Tables

1 Introduction

Speech production is a complex motor activity that requires the precise synchronization of respiration, phonation, and articulation. Beyond merely conveying words (linguistic content), this coordinated process simultaneously embeds a wealth of extra-linguistic information: the *how* rather than the *what* of speech. This layer constitutes the field of paralinguistics, rooted in the Greek *para* ("alongside"). Paralinguistic cues including- intonation, rhythm, loudness, timbre, and voice-quality modulate the verbal message, allowing listeners to infer critical information about the speaker's emotional state, attitude, accent, gender, personality, or health condition, making them indispensable bridges between linguistic intent and affective, physiological, and social context.

Paralinguistic phenomena are categorized along two interacting dimensions that exist on a temporal continuum: States and Traits *(Schuller et al., 2013a)*. States are transient, situational variations (e.g., emotion, stress, cognitive load) that reflect the speaker's short-term affective or cognitive condition. Conversely, Traits represent relatively stable and long-term characteristics (e.g., age, gender, personality, pathology) rooted in physiology or habitual behavior. This continuum is critical because stable traits fundamentally constrain and filter how fleeting states are acoustically expressed.

Foundational work in phonetics and psychology established links between measurable acoustic correlates (such as pitch, loudness, tempo, and breathiness) and human attributes like emotion and personality *(K. R. Scherer, 1978)*. With the advent of speech signal processing, these observations evolved into computational models aiming to infer speaker states and traits directly from acoustic evidence, marking the emergence of computational paralinguistics *(Schuller and Batliner, 2013)*. The modeling approach in this field has since dramatically evolved from traditional systems relying on descriptive phonetics and handcrafted acoustic features (like pitch statistics and Mel-Frequency Cepstral Coefficients) *(Schuller, 2018)* to modern deep learning techniques *(LeCun et al., 2015)*. This progression allowed convolutional and recurrent networks to learn hierarchical neural representations directly from audio *(Hinton et al., 2012; Palaz et al., 2019)*. Most recently, self-supervised representation learning has become the state of the art, with Speech Foundation Models (SFMs) *(A. Mohamed et al.,*

2022) such as wav2vec 2.0 *(Baevski et al., 2020)* and its successors *(Conneau et al., 2021; W.-N. Hsu et al., 2021; Yang et al., 2021; S. Chen et al., 2022)* offering generalizable, data-driven neural embeddings that implicitly encode a vast array of phonetic, prosodic, and affective information. This shift from knowledge-based feature engineering to learned, data-driven representations defines the current technical landscape.

While data-driven neural representations have consistently outperformed handcrafted acoustic features across paralinguistic tasks, they often remain opaque and difficult to interpret *(Favaro et al., 2023)*. This lack of transparency poses a significant challenge as paralinguistic speech technologies transition from controlled laboratory settings to real-world applications. In other words, as these technologies are tightly coupled with end users, the demand for explainability is no longer a theoretical requirement but a practical necessity. For example, in healthcare domains, transparent and interpretable models are more desirable than those offering marginal performance gains without explainability. To advance robust paralinguistic analysis, this thesis evaluates the efficacy and limitations of modern neural representations. It systematically investigates the evolving spectrum of speech modeling, ranging from handcrafted acoustic features to data-driven neural embeddings by addressing the following research questions (RQs):

RQ1. Whether emotion discriminative information be effectively learned/modeled from sub-word level short segment of speech (of duration around 250 ms)?

RQ2. What is the emotion discriminative capacity of phonetically aware neural representations?

RQ3. How can SFMs, pre-trained on healthy speech, be effectively utilized and adapted for low-resource pathological speech?

RQ4. How robust are the neural representations in encoding states (like depression) within the context of governing traits (like Parkinson's Disease)?

While the core work centers on learned representations, we consistently retain knowledge-based features for comparison, a step that ensures the neural model performance is grounded in established acoustic-phonetic evidence. Through this progression, our investigation examines how different speech representations allow us to analyze the way transient states (e.g., emotions) and enduring traits (e.g., Parkinson's disease) are both encoded and inferred from the acoustic signal.

Thesis Outline and Contributions

The thesis is structured as following.

Chapter 2: Background
This chapter provides a background necessary to contextualize and understand the subsequent studies presented in this thesis. It reviews key concepts, frameworks, and prior research in speech processing, paralinguistics, and deep learning.

Chapter 3: Learning emotion information from short segments of speech
This chapter investigates **RQ1**. The conventional approach to Speech Emotion Recognition (SER) typically involves mapping variable-length, frame-level acoustic descriptors into a single fixed-dimensional feature vector. More recent methods model utterance-level representations-such as Mel-Frequency Cepstral Coefficients (MFCCs) or longer duration raw-audio inputs, using deep neural networks for SER. In contrast, this study proposes an end-to-end Convolutional Neural Network (CNN) framework and demonstrates that emotional information can be effectively captured within very short speech segments ($\approx$ 250 milliseconds). This short-segment approach is validated across diverse paralinguistic tasks. This research work has been presented in the following publications:

Purohit et al. (2023b): **Tilak Purohit**, Sarthak Yadav, Bogdan Vlasenko, S Pavankumar Dubagunta, and Mathew Magimai Doss. (2023). Towards Learning Emotion Information from Short Segments of Speech. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*.

Purohit et al. (2022): **Tilak Purohit**, 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 ICML Expressive Vocalizations (ExVo) Workshop and Competition (ICML-workshop 2022)*. [Ranked among the top 5 submissions]

Yadav et al. (2022): Sarthak Yadav, **Tilak Purohit**, Zohreh Mostaani, Bogdan Vlasenko, and Mathew Magimai Doss. (2022). Comparing biosignal and acoustic feature representation for continuous emotion recognition. In *Proceedings of ACM International Conference on Multimedia: 3rd MuSe Workshop and Challenge (ACM 2022)*.

Chapter 4: Phonetically aware neural representations for speech emotion recognition
This chapter focuses on **RQ2**, and explores how emotion information manifests through subtle phonetic variations encoded within speech. It proposes a transcription-free framework for modeling phonetic information by leveraging embeddings from networks pretrained on phoneme or grapheme recognition tasks. Evaluated across multiple benchmark emotion corpora, the proposed phonetically aware neural representation consistently outperforms traditional handcrafted acoustic features in SER. This research work has been presented in the publication:

Purohit et al. (2023a): **Tilak Purohit**, Bogdan Vlasenko, and Mathew Magimai Doss. (2023). Implicit phonetic information modeling for speech emotion recognition. In *Proceedings of*

Interspeech, 2023.

Chapter 5: Probing speech foundation models for emotion information recovery
This chapter extends and probes the findings of Chapter 4, specifically the trade-off between optimizing for linguistic content (Automatic speech recognition or ASR) and preserving paralinguistic sensitivity (SER) for SFM. It introduces a framework to systematically investigate whether the paralinguistic information lost during task-specific ASR fine-tuning in SFMs can be recovered, and its consequences to the networks latent representations. This research work has been presented in the publication:

Purohit and Magimai-Doss (2025): **Tilak Purohit** and Mathew Magimai Doss. (2025). Emotion information recovery potential of wav2vec2 network fine-tuned for speech recognition task. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*.

Chapter 6: Model adaptation for Parkinson's Disease detection from speech
This chapter investigates **RQ3**, the adaptation strategies for low-resource pathological speech, with a focus on Parkinson's Disease (PD). In the framework of SFMs, it presents one of the first systematic comparison of approaches such as layer selection, full fine-tuning, and parameter-efficient adaptation (LoRA), highlighting their respective trade-offs in deriving discriminative features. This research work has been presented in the publication:

Purohit et al. (2025a): **Tilak Purohit**, Barbara Ruvolo, Juan Rafael Orozco-Arroyave, and Mathew Magimai Doss. (2025). Automatic Parkinson's disease detection from speech: Layer selection vs adaptation of foundation models. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*.

Chapter 7: Speech-based analysis of depression comorbidity in Parkinson's Disease
Building on the insights from the previous chapters, this chapter investigates **RQ4**. The modeling of an affective state depression within a neurological trait PD. To the best of our knowledge, in a first-of-its-kind study, it examines both interpretable handcrafted acoustic features and non-interpretable SFM-derived representations for this challenging task, where overlapping affective and neurological factors complicate inference.

Purohit et al. (2025b): **Tilak Purohit**, Barbara Ruvolo, Juan Rafael Orozco-Arroyave, and Mathew Magimai-Doss. (2025). On detection of depression in parkinson's disease patients' speech: Handcrafted features vs. speech foundation models. In *Automatic Assessment of Parkinsonian Speech, Communications in Computer and Information Science. Springer Nature Switzerland AG (CCIS 2025)*

Chapter 8: Conclusions and Future Directions
Concludes the thesis with suggestions for future research directions.

2 Background

2.1 Introduction to paralinguistics

Speech is the most fundamental mode of human communication, a multi-layered signal carrying linguistic information (the *what*) inextricably bound to extra-linguistic information (the *how*). The study of this extra-linguistic dimension, the vocal features that accompany and modulate the verbal message constitutes the field of paralinguistics *(Schuller et al., 2013a)*. Rooted in the Greek prefix *para-* (alongside), paralinguistics encompasses a vast array of vocal phenomena including intonation, pitch variation, loudness, speech rate, vocal texture, and silence patterns. These cues serve as indispensable channels for conveying information related to the speaker's internal state, identity, and physiological condition, thereby providing crucial context for the linguistic content.

The significance of paralinguistic analysis stems from its capacity to offer a window into non-verbal, often involuntary, human characteristics and reactions. Unlike explicit linguistic content, paralinguistic cues often reveal a speaker's genuine emotional state or physiological condition. This makes them particularly valuable for applications such as human-computer interaction, intelligent tutoring systems, clinical diagnostics, and mental-health monitoring *(Brederoo et al., 2021; K. Feng and Chaspari, 2024; Kothare et al., 2025)*. Research in this domain is highly interdisciplinary, drawing heavily from acoustic phonetics, psychology, signal processing and machine learning *(Schuller and Batliner, 2013)*. The central ambition of paralinguistic community is to establish a robust framework for understanding, extracting, and modeling these subtle vocal modulations. This research is fundamentally guided by the principle: **'Interested more in *how* you say, than *what* you say.'**

The remainder of this chapter is structured as follows. Section 2.2 introduces the paralinguistic state-trait continuum, outlining the distinction between transient speaker states and stable speaker traits that form the conceptual foundation of paralinguistic modeling. Section 2.3 presents various emotion models and labeling paradigms, contrasting categorical and dimensional representations and discussing their implications for emotion annotation. Section 2.4 then explain the notion of typical and atypical speech. Section 2.5 reviews the progression of

paralinguistic feature extraction and representation, from handcrafted acoustic descriptors to modern deep representation learning. Section 2.6 introduces recent advances in Speech Foundation Models (SFMs). Finally, Sections 2.7 and 2.8 describe the objective functions and evaluation metrics employed in paralinguistic research, completing the conceptual and methodological groundwork for the subsequent chapters.

2.2 The paralinguistic state-trait continuum

A central concept in paralinguistic analysis is the distinction between *states* and *traits*, which exist along a dynamic temporal continuum *(Schuller and Batliner, 2013)*. This continuum is essential because stable traits fundamentally constrain and filter how transient states are acoustically expressed, leading to complex interactions that influence modeling accuracy. Understanding this dichotomy is important for designing robust and generalizable paralinguistic systems.

Paralinguistic states: Speaker states are defined as short-term, transient variations in speech characteristics that reflect the speaker's immediate affective, cognitive, or physiological condition. These are highly dynamic, contextual features that fluctuate rapidly based on environmental stimuli or internal processes, and their scope extends beyond discrete emotions to include factors like cognitive load and psychological stress. Specific examples of these states encompass emotion-related affects such as stress, confidence, uncertainty, frustration, and pain; however, the most intensely studied paralinguistic state remains emotions (e.g., anger, joy, sadness, fear) *(Schuller, 2018)*.

Paralinguistic traits: In contrast to transient states, speaker traits represent relatively stable, long-term characteristics rooted in the speaker's fixed physiology, ingrained behavioral habits, or established medical conditions. These characteristics are largely independent of the momentary communication context. Gender and age are primary biometric traits. Permanent pathology such as Parkinson's Disease or Alzheimer Disease can be considered a trait.

2.3 Emotion models and Labeling paradigms

Theoretical models of emotion provide the foundation upon which computational frameworks are built, defining both the structure of the emotional space and the strategies used to annotate data. This section first outlines the principal emotion modeling approaches, categorical and dimensional that describe affect either as discrete classes or as points in a continuous space. It then discusses the corresponding labeling paradigms, which determine how emotional states are annotated and interpreted in speech corpora.

2.3.1 Categorical versus Continuous

The systematic study of emotion began with Ekman's foundational *(Ekman, 1971)* work in the 1970s, which established one of the earliest systematic frameworks for understanding human affective states. Since then, diverse fields have contributed to various interpretations and models of emotion such as: Grandjean et al., 2008; Tracy and Randles, 2011; Gunes and Schuller, 2013; Marsella and Gratch, 2014. While a universally accepted model remains elusive, two principal theoretical paradigms: categorical (discrete) and dimensional (continuous) form the cornerstone of emotion representation in computational paralinguistics.

The categorical emotion theory, primarily associated with Ekman *(Ekman, 1971; Ekman et al., 1999)*, posits that emotion consists of a small set of fundamental, discrete emotions commonly-anger, disgust, happiness, sadness, fear, and surprise. These are considered innate, universally expressed across cultures, reliably elicited, and consistently signaled through facial, vocal, or physiological cues. Though over ninety distinct definitions of basic emotions have been suggested *(Plutchik, 2001; Plutchik, 2003; Gunes et al., 2011)*, the core concept is that these are distinct, non-overlapping states. However, the idea of strict universality is debated, as cultural and contextual differences can modify both the expression and perception of emotions. Furthermore, discrete emotions can combine to form complex or compound affective states *(Ekman et al., 1999)*. A prominent extension is Plutchik's wheel of emotions *(Drews and Krohn, 2007)* based on *(Plutchik, 1982)*, which organizes eight primary emotions (e.g., happiness, sadness, anger, fear) in a circumplex model as opposing pairs. Secondary and tertiary emotion combinations, or dyads (e.g., anticipation + trust = hope; anger + disgust = contempt), arise from these primaries. This hierarchical, combinatorial view also incorporates intensity variation (e.g., serenity → joy → ecstasy), emphasizing that affective experiences exist along a continuum of blended states rather than just fixed categories.

In contrast to categorical models, the dimensional emotion theory represents emotions as coordinates within a continuous, multidimensional affective space, allowing for a richer, more flexible basis for modeling emotional variation in continuous speech. Key models include Russell's circumplex model of affect *(Russell, 1980)* and Mehrabian's pleasure-arousal-dominance model *(Mehrabian, 1996)*. In these frameworks, affect is characterized by continuous axes: Valence (or pleasure) represents emotional polarity, measuring the magnitude of human feeling from extreme ecstasy (positive) to distress (negative). Arousal (or activation) captures the intensity or energy level of the emotion, spanning from highly excited or energized states to calm or quiescent states. The third axis, Dominance (or control), describes the perceived sense of influence over the surrounding environment, ranging from feeling dominant and in control to feeling weak or powerless. Together, these axes define a continuous emotion manifold where discrete emotions can be positioned and compared quantitatively; for example, anger is empirically characterized in the Valence, Arousal, Dominance (VAD) space by an average of negative valence, high arousal, and strong dominance *(Grimm et al., 2006)*. This quantitative mapping serves as a bridge between discrete emotion labels and continuous dimensions, enabling finer granularity in automatic modeling. While three dimensions provide a com-

prehensive affective space, it is often claimed that the two core dimensions of Valence and Arousal are sufficient to represent the vast majority of different human emotional experiences *(Russell and Mehrabian, 1977)*.

2.3.2 Annotation and Labeling paradigms

The choice of modelling approach (categorical or dimensional) directly dictates the annotation protocol. Emotion annotation can follow either categorical or dimensional paradigms. In the categorical approach, emotions are labeled as discrete classes such as anger, happiness, sadness, or neutral, reflecting prototypical affective states. In contrast, the dimensional approach represents affect in a continuous space—most commonly the valence-arousal (V-A) model, where valence indicates emotional polarity (pleasant-unpleasant) and arousal reflects activation intensity. While categorical labels offer interpretability and simplicity for classification tasks, they often fail to capture mixed or subtle affective expressions. Dimensional labels, though better suited for describing emotional dynamics, introduce challenges related to subjectivity, annotation consistency, and inter-rater reliability, as different annotators may perceive emotional intensity differently. Consequently, large-scale corpora such as the MSP-Podcast *(Lotfian and Busso, 2017)* employ strategies like continuous annotations, annotator normalization, and gold-standard fusion to mitigate these inconsistencies and achieve reliable affective labeling *(Busso et al., 2008; Lotfian and Busso, 2017)*. In categorical labeling, annotators (often lay listeners) are typically asked to select the single most appropriate emotional label from a fixed list for each utterance *(Steidl et al., 2009)*. The simplicity of this approach is offset by its major drawback: it forces subtle emotional reality into mutually exclusive categories. This oversimplification often reduces inter-annotator agreement and prevents the labels from fully capturing the emotional complexity. Dimensional labeling, on the other hand, requires annotators to rate each utterance along continuous scales (e.g., 1-7 or -3 to +3) for valence and arousal, sometimes including a third dimension such as dominance. This process is carried out using graphical tools, and the resulting scores from multiple annotators are averaged or fused to form a consensus representation.

2.4 Typical and Atypical speech

In computational paralinguistics, the objective is to identify, describe, and model variations in speech that reveal behavioral, affective, or physiological phenomena. These phenomena may be *typical*: frequent and broadly representative of general speaking behavior or *atypical*: occurring less frequently due to specific contexts, traits, or pathologies *(Schuller et al., 2013a)*. The task of paralinguistic analysis thus involves not only feature extraction and modeling, but also a systematic comparison between typical and atypical manifestations of speech.

In clinical and developmental contexts, the term “atypical” carries yet another meaning, often referring to deviations from normative patterns of speech and language development. Instead of pathologizing, this terminology provides a neutral framework for describing diversity in

communication abilities. For instance, when examining pathological conditions, studies commonly draw a sharp distinction between "typically developing" and "atypically developing" speakers *(Janbakhshi, 2022)*.

Irrespective of terminological nuance, the scientific motivation remains consistent: to identify, characterize, and improve the processing of speech that diverges from conventional models. Research in this area often seeks to (i) uncover the acoustic and linguistic differences between typical and atypical speech, (ii) enhance the robustness of algorithms that fail under atypical conditions, or (iii) enable diagnostic and monitoring systems capable of detecting atypical speech patterns linked to medical or developmental conditions. The broader ambition of paralinguistic research: to understand the variability of human speech across physiological, developmental, and affective dimensions, and to design computational systems that remain sensitive and robust across both typical and atypical domains. In the context of this thesis, the notion of atypical speech becomes particularly relevant in studying pathological speech conditions such as Parkinson's disease, where deviations in articulation, prosody, and phonation constitute salient but underrepresented paralinguistic cues.

2.5 Paralinguistic feature extraction and representation

The history of paralinguistic modeling can be charted by the evolution of the features used to represent the acoustic signal, moving from hand-engineered, theoretically motivated statistics to data-driven, automatically learned representations *(Schuller, 2018)*.

2.5.1 Handcrafted feature engineering

In the pre-deep learning era, the foundation of paralinguistic analysis rested on extracting interpretable, low-level acoustic descriptors (LLDs) and aggregating them into utterance-level summary statistics. This approach is highly effective because it directly ties computational features to established phonetic and psychological theories *(Schuller and Batliner, 2013)*.

2.5.1.1 Low-Level Descriptors (LLDs)

LLDs are computed on short, fixed-length frames of speech (typically 10-50 ms) and fall into three primary categories *(Eyben, 2015)*:
(1) Acoustic-Prosodic Features: These capture the suprasegmental elements of speech. They include fundamental frequency (F_0) and its variants (pitch, logarithmic F_0 difference), energy (e.g., RMS energy, loudness), and temporal features (e.g., speaking rate, duration of voiced/unvoiced segments, pause count and duration).
(2) Spectral Features: These represent the short-term distribution of energy across the frequency spectrum, relating to timbre and voice quality. The most common are Mel-Frequency Cepstral Coefficients (MFCCs), which emulate human auditory perception, and their deriva-

tives (Δ and $\Delta\Delta$ coefficients) *(Logan et al., 2000)*. Others include spectral moments (centroid, spread, skewness, kurtosis) and subband energy ratios.
(3) Voice Quality Features: These capture characteristics of the vocal source (vibrations of the vocal folds) and are highly indicative of stress, emotion, and pathology. Examples include jitter (perturbation of F_0), shimmer (perturbation of amplitude), and features related to the glottal source (e.g., harmonic-to-noise ratio, spectral tilt) *(Alku et al., 2005)*.

2.5.1.2 Suprasegmental features

Unlike tasks such as Automatic Speech Recognition (ASR), which deal with short-term acoustic phenomena like phonemes, the conventional approach for modeling emotion recognition and speaker states for over a decade, relies on mapping variable-length, frame-level LLDs into a single, fixed-dimensional feature vector suitable for standard classification or regression models. Two major categories of mechanisms achieve this fixed-length feature vector representation from variable-length segments:

(1) Statistical Functionals: This approach converts time-series acoustic data into a static vector by applying various statistical functionals (e.g., mean, median, standard deviation, minimum, maximum, quartiles, and linear regression coefficients) across the entire utterance *(Schuller and Batliner, 2013; Schuller et al., 2013b)*. This process collapses the continuous Low-Level Descriptors (LLDs) into a single, high-dimensional vector that comprehensively describes the acoustic properties of the utterance. Early efforts, such as the INTERSPEECH Computational Paralinguistics ChallengE (ComParE) feature sets, defined extensive standards comprising over 6,000 features *(Schuller et al., 2013b)*. This led to the development of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) *(Eyben et al., 2016)*, which is now a highly recommended standard for paralinguistic analysis. eGeMAPS offers a computationally simpler, parsimonious, and psychologically relevant set of 88 LLDs and their functionals.

(2) Bag-of-Audio-Words (BoAW): An alternative fixed-length representation is achieved using the Bag-of-Audio-Words (BoAW) approach, often implemented via the openXBOW toolkit *(Schmitt and Schuller, 2017)*. This method works by first quantizing the frame-level acoustic LLDs based on a pre-learned codebook. Each audio segment is then represented as a histogram of these resulting “acoustic words," creating a feature vector whose length is fixed by the size of the codebook. This approach has proven highly effective across various speech applications, including speech-based emotion recognition and acoustic event detection *(Lim et al., 2015; Schmitt et al., 2016)*.

2.5.2 Classical Machine Learning (ML) classifiers

Based on the extraction of features as outlined in the previous section, the next step involves designing a system capable of mapping these features to the target variable of interest—such as

a speaker's age, emotional state, or health condition. This process constitutes the classification stage, where the goal is to learn a discriminative function that separates samples belonging to different classes or predicts continuous attributes in the case of regression. The following classical machine learning classifiers were used in the experiments:

(1) Support Vector Machines (SVMs): SVMs classify data by projecting input features into a higher-dimensional space where the separation between classes can be represented by a hyperplane. The objective is to determine the optimal hyperplane that maximizes the margin between positive and negative samples. Originally formulated as the maximum margin classifier *(Vapnik and Lerner, 1963)*, the approach was later extended to the soft-margin SVM to handle non-separable data by introducing slack variables *(Cortes and Vapnik, 1995)*. The subsequent integration of kernel functions enabled the construction of non-linear decision boundaries in the transformed feature space *(Boser et al., 1992)*. Over time, the framework was further generalized to support multi-class classification tasks *(C.-W. Hsu and Lin, 2002)*.

(2) Decision Trees and Random Forests: Decision trees classify data by recursively splitting it according to feature thresholds that optimize criteria such as information gain or Gini impurity. Each internal node represents a decision rule, while the terminal nodes correspond to class labels. Extending this idea, Random Forests (RF) aggregate multiple decision trees trained on random subsets of the training data and feature dimensions, thereby reducing variance and improving generalization performance. The ensemble prediction is obtained by majority voting among the constituent trees *(Breiman, 2001)*.

(3) Multilayer Perceptrons (MLPs): These are among the earliest forms of feed-forward neural networks *(Rosenblatt, 1958)*, composed of multiple layers of interconnected neurons. Each neuron performs a linear transformation of its inputs followed by a non-linear activation function, allowing the network to model complex, non-linear relationships between input features and output labels. The parameters (weights and biases) are optimized via backpropagation and gradient descent to minimize a task-specific loss function *(Rumelhart et al., 1986)*. MLPs can approximate any continuous function given sufficient hidden units and layers, making them a foundational building block for modern deep learning architectures *(Goodfellow et al., 2016)*.

2.5.3 Neural representation learning models

The transition to deep learning architecture marks a paradigm shift, enabling models to learn optimal feature representations directly from the raw speech signal or a simple spectrogram, eliminating the need for feature engineering *(Trigeorgis et al., 2016; Tzirakis et al., 2017)*.

2.5.3.1 Recurrent Neural Networks (RNNs)

RNNs *(Elman, 1990)*, particularly their variant like Long Short-Term Memory (LSTM) *(Hochreiter and Schmidhuber, 1997)*, are naturally suited for sequential data like speech *(Graves et*

al., 2013). LSTMs excel at capturing long-range dependencies across the utterance, crucial for prosodic patterns which span several seconds (e.g., the overall contour of sadness or a sustained vocal tremor) *(Trigeorgis et al., 2016)*. They process the frame-level input (typically MFCCs or filter banks) sequentially, aggregating temporal context before feeding the final hidden state to a classification or regression layer.

2.5.3.2 Convolutional Neural Networks (CNNs)

CNNs, originally designed for image processing *(LeCun et al., 1989; LeCun and Bengio, 1998)*, are highly effective in paralinguistics when applied to 2D representations of speech, such as spectrograms or Mel-Spectrograms *(Yu et al., 2013)*. Spectrograms transform the 1D time-domain signal into a 2D time-frequency image, where local patterns (e.g., formants, rapid pitch changes, noise) become visually apparent. CNN kernels automatically learn local, translation-invariant patterns in this representation, capturing subtle phonetic and acoustic details that characterize specific emotions or traits *(Z. Huang et al., 2014; Amiriparian et al., 2017)*.

2.5.3.3 Attention Mechanism

Modern deep learning architectures frequently incorporate attention mechanisms *(Bahdanau, 2014)*, which enable models to dynamically assign importance to different parts of the input sequence rather than processing all elements with equal weight. In the context of speech processing, attention helps the network focus on emotionally or phonetically salient regions, such as: stressed syllables, pitch variations, or high-energy vocal bursts, that are most informative for the downstream task (e.g., detecting anger or stress) *(Mirsamadi et al., 2017)*.

The fundamental formulation, known as Scaled Dot-Product Attention, computes a similarity score between a query vector (Q) and a set of key vectors (K), and uses these scores to derive a weighted sum over the corresponding value vectors (V):

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, Q, K, and V are projections of the input features, and d_k denotes the dimensionality of the key vectors. The Softmax function normalizes the attention weights, ensuring they form a probability distribution over the sequence elements. This mechanism can be interpreted as a form of soft alignment, allowing the model to decide which time frames or spectral regions to prioritize during training.

2.5.3.4 Transformer Encoder

The Transformer *(Vaswani et al., 2017)* architecture, originally proposed for machine translation, has since become a dominant framework for sequence modeling in speech and language

tasks. It replaces the recurrent operations of earlier models with parallelizable self-attention layers, enabling efficient modeling of long-range dependencies, an essential property for capturing prosodic, contextual, and affective patterns in speech *(Ramet et al., 2018)*.

A single Transformer encoder layer transforms an input sequence of embeddings $X \in \mathbb{R}^{T \times d}$ (where T is the sequence length and d is the embedding dimension) through two main sub-layers:

$$Z = \text{LayerNorm}\big(X + \text{MultiHeadAttention}(X)\big)$$
$$Y = \text{LayerNorm}\big(Z + \text{FeedForward}(Z)\big)$$

The first sub-layer employs multi-head self-attention to aggregate contextual information across all time steps, while the second applies a position-wise feed-forward network to refine these representations. Residual connections and layer normalization ensure stable training and facilitate gradient flow.

In the speech domain, Transformer encoders serve as the backbone of several Speech Foundation Models, such as WAV2VEC2.0, HUBERT, and WAVLM.

2.6 Speech Foundation Models (SFMs)

In recent years, the emergence of foundation models has marked a major shift in the landscape of representation learning *(Bommasani et al., 2021)*. These models are characterized by large-scale training on diverse and extensive datasets, enabling them to generalize across a broad spectrum of downstream tasks through adaptation techniques such as fine-tuning. Among the various training paradigms, self-supervised learning (SSL) has become a dominant strategy for developing such models, as it allows the extraction of informative representations from unlabeled data. The overall framework typically involves two distinct stages: during the pretraining phase, a general-purpose model, often referred to as the upstream model, is trained using SSL objectives to capture high-level representations of the input domain. Subsequently, in the downstream phase, this pretrained model is either fine-tuned end-to-end or used as a frozen feature extractor within a supervised learning setup to address specific target tasks. In the speech research community, SFMs have been widely adopted for a range of paralinguistic tasks *(Yang et al., 2021; A. Mohamed et al., 2022; Zhang et al., 2024)*. The three most prominent SFMs used in this study are outlined below.

1. Wav2Vec 2.0 (W2V2) *(Baevski et al., 2020)*: W2V2 integrates masking and contrastive learning for self-supervised speech representation learning. The network takes raw waveforms as input and first encodes them into latent representations using a convolutional feature encoder. Spans of these latent representations are then randomly masked before being passed through a Transformer-based context network that generates contextualized embeddings.

Training is performed via a contrastive pretext task, where the model learns to distinguish

the true latent representation from a set of distractors. Specifically, given the context network output c_t corresponding to a masked time step t, the model is trained to identify the correct quantized latent representation q_t within a candidate set Q_t of size $K+1$, consisting of q_t and K distractors. Distractors are uniformly sampled from other masked time steps within the same utterance. The contrastive loss is defined as:

$$\mathscr{L}_m = -\log \frac{\exp\left(\text{sim}(c_t, q_t)/\kappa\right)}{\sum_{\tilde{q}\in Q_t} \exp\left(\text{sim}(\mathbf{c}_t, \tilde{q})/\kappa\right)},$$

where $\text{sim}(c_t, q_t)$ denotes the cosine similarity between the context representation and the quantized latent speech representation, and κ is a temperature parameter.

The W2V2 models are pretrained on large-scale audio corpora such as LibriSpeech(960 hours) *(Panayotov et al., 2015)* and LibriVox (60k hours) *(Kearns, 2014)*, both derived from audiobook recordings.

2. HuBERT: Instead of employing sophisticated self-supervised methods for discretizing continuous speech signals, HuBERT (Hidden-Unit BERT) *(W.-N. Hsu et al., 2021)* investigates the effectiveness of a simpler approach, using k-means clustering on MFCC features to create discrete target units. During training, the model predicts the pre-assigned k-means cluster labels for masked segments of continuous speech representations. Like Wav2Vec 2.0, HuBERT processes raw audio through a convolutional encoder followed by a Transformer network, with masking applied between these stages. However, unlike contrastive learning frameworks that rely on negative sampling to prevent representational collapse, HuBERT directly optimizes a cross-entropy objective between the predicted and true cluster identities. The model is pretrained on large-scale speech datasets such as LibriSpeech (960 hours) and Libri-Light (60k hours) *(Panayotov et al., 2015)*.

3. WavLM: Building upon the HuBERT architecture, WavLM *(S. Chen et al., 2022)* focuses on jointly modeling linguistic content and speaker characteristics. It introduces a content-aware gated relative position bias within the Transformer's self-attention mechanism, enhancing recognition performance beyond that of the convolution-based front ends in HuBERT and Wav2Vec 2.0. Furthermore, WavLM incorporates an utterance-mixing strategy, where overlapping speech from multiple speakers is artificially generated to enrich the pretraining data. The model learns to predict the masked regions corresponding to the primary speaker, thereby improving its sensitivity to both linguistic and paralinguistic cues (e.g., speaker identity and style). Trained on an extended 94k-hour dataset that combines the corpora used for HuBERT and Wav2Vec 2.0, WavLM achieves strong results on multi-speaker tasks such as speech separation and diarization.

2.7 Objective functions

The objective/loss function is determined by the labeling paradigm, with classification tasks utilizing Cross-Entropy and regression tasks employing Mean Squared Error (MSE) or Concordance Correlation Coefficient (CCC) *(Ringeval et al., 2015a)*.

1. Categorical model (Classification loss): For a categorical task with C classes (e.g., angry, joy, sadness), the standard loss function is the Categorical Cross-Entropy Loss ($\mathscr{L}_{\text{CE}}$):

$$\mathscr{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c})$$

where N is the number of samples, $y_{i,c}$ is a binary indicator (0 or 1) if class c is the correct classification for observation i, and $\hat{y}_{i,c}$ is the model-predicted probability of observation i belonging to class c.

2. Continuous model (Regression loss): For continuous state/trait prediction (e.g. depression severity-score, age), the Mean Squared Error ($\mathscr{L}_{\text{MSE}}$) is a common objective:

$$\mathscr{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where N is the number of samples, y_i is the ground-truth continuous label (e.g., arousal score), and $\hat{y}_i$ is the model's predicted score.

For emotion regression tasks (specifically Valence and Arousal prediction), the Concordance Correlation Coefficient (CCC) *(Nickerson, 1997)* is used to measure the agreement between the network's predicted values and the human-annotated gold-standard scores *(Ringeval et al., 2015a; Ringeval et al., 2015b)*. CCC is an effective metric because it quantifies both accuracy (how close the means are) and precision (how close individual observations are). It combines the Pearson Correlation Coefficient (PCC), which measures linear association, and a component that corrects for mean and scale differences. CCC values range from -1 to 1, where 1 signifies perfect agreement and 0 means no agreement. The CCC loss function is formulated as:

$$\mathscr{L}_{\text{CCC}} = 1 - \text{CCC},$$

where the CCC is defined as:

$$\text{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}.$$

Here, μ_x and μ_y denote the means of the predicted values $\hat{y}$ and the ground-truth labels y, respectively, while σ_x and σ_y represent their corresponding standard deviations. The term ρ refers to the Pearson Correlation Coefficient (PCC) between $\hat{y}$ and y, which measures

the strength and direction of their linear relationship. A value of ρ close to 1 indicates a strong positive correlation, while values near -1 and 0 indicate strong negative and no linear correlation, respectively.

2.8 Evaluation Metrics

The performance of the proposed systems are assessed using a range of evaluation metrics, selected according to the task formulation- classification or regression, and the application domain, spanning state (SER) and trait (PD speech detection) task. This section outlines the metrics employed throughout the thesis.

A classification model can either correctly classify a sample in its actual class, or incorrectly predict it to belong to another class. By comparing the predicted class with the ground truth, one can obtain four possible outcomes: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Based on these, several evaluation metrics can be derived, as described below.

- **Accuracy (ACC):** The proportion of correctly classified samples over the total number of samples. This metric provides a general measure of the model's overall performance.

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision (P):** The ratio of true positive predictions to the total number of positive predictions made by the model. It reflects the model's ability to avoid false positives.

$$\text{P} = \frac{TP}{TP + FP}$$

- **Recall (R) / Sensitivity:** Also known as the True Positive Rate (TPR), recall is the ratio of correctly identified positive instances to the total number of actual positive instances. It indicates how effectively the model identifies all relevant cases.

$$\text{R} = \frac{TP}{TP + FN}$$

- **Specificity:** Also known as the True Negative Rate (TNR), specificity measures the proportion of correctly identified negative (control) samples. It complements sensitivity and is particularly important in pathological or diagnostic tasks.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

- **Unweighted Average Recall (UAR):** For multi-class classification, UAR is computed as the mean of class-wise recall scores, providing a balanced estimate of performance

under class imbalance.

$$\text{UAR} = \frac{1}{C}\sum_{i=1}^{C}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$$

where C is the total number of classes, and TP_i, FN_i are the true positive and false negative counts for class i.

- **Concordance Correlation Coefficient (CCC):** For regression-based tasks such as continuous emotion prediction, CCC evaluates the agreement between predicted and ground-truth values by combining both accuracy and precision. It is defined as:

$$\text{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

where ρ is the Pearson Correlation Coefficient between predictions x and ground-truth values y, σ_x and σ_y denote their standard deviations, and μ_x and μ_y represent their mean values. A CCC of 1 indicates perfect agreement, 0 indicates no agreement, and -1 represents complete disagreement.

- **Mean Squared Error (MSE):** MSE quantifies the average squared difference between the predicted and target values, penalizing larger errors more heavily.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- **Mean Absolute Error (MAE):** MAE measures the average magnitude of errors without considering their direction, offering an interpretable measure of overall prediction deviation.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Together, these metrics provide a comprehensive evaluation framework that spans categorical, continuous model for state and trait speech analysis tasks.

2.9 Summary

This chapter outlined the key concepts of computational paralinguistics, emphasizing how speech conveys information beyond words through vocal cues linked to emotional, cognitive, and physiological states. It introduced the state-trait continuum, distinguishing transient affective states from stable speaker traits. The chapter reviewed major emotion models and labeling paradigms (categorical vs. dimensional), discussed typical and atypical speech variations, and traced the evolution from handcrafted acoustic features to deep and self-supervised speech foundation models. Finally, it summarized the key loss functions and evaluation metrics used for both categorical and continuous paralinguistic tasks, establishing the groundwork for the subsequent chapters.

3 Learning emotion information from short segments of speech

In speech paralinguistics, conventional approaches have predominantly relied on suprasegmental modeling *(Schuller and Batliner, 2013)*. In this paradigm, low-level acoustic descriptors computed on short frames are transformed into fixed-length representations using statistical functionals (e.g., mean, variance, percentiles, regression coefficients), effectively summarizing how acoustic parameters evolve over time. Such utterance-level representations have demonstrated strong performance in capturing broad speaker states and traits- such as emotion, gender, or personality. Recent approaches to speech emotion recognition (SER) predominantly adopt an utterance-level modeling framework, where affective information is derived from longer speech segments. Some studies extract statistical or spectral representations, such as Mel-Frequency Cepstral Coefficients (MFCCs), and input them to traditional classifiers (e.g., SVMs) or deep architectures *(Ghosh et al., 2016; Xia and Y. Liu, 2016; Neumann and T. Vu, 2017; Kim and Shin, 2019; Zhao et al., 2019; Peng et al., 2021)*. Others bypass handcrafted features altogether, modeling long-duration raw audio signals (typically spanning 4-6 seconds) directly through end-to-end neural architectures such as CNN-LSTMs and TDNNs *(Zhao et al., 2019; J.-L. Li et al., 2020; Kumawat and Routray, 2021)*.

However, the aforementioned methods overlooks fine-grained temporal and spectral variations that may themselves carry rich emotional cues. Emotional expression is conveyed not only by extended prosodic contours but also by rapid localized fluctuations in pitch, energy, and spectral balance, often within the span of a single syllable *(C. M. Lee et al., 2004; Vlasenko and Wendemuth, 2013; Dhamyal et al., 2020; Yuan et al., 2021)*. Aggregating speech over long windows can average out these instantaneous modulations, erasing discriminative cues embedded in the signal's microstructure.

Motivated by these insights, we investigate whether emotion-relevant information can be learned directly from short, speech segments, approximately the length of a syllable (≈250 ms). Evidence from recent paralinguistic studies, such as the ACM Multimodal Sentiment Analysis Challenge (MuSE) *(Stappen et al., 2021; Amiriparian et al., 2022)* and ICML Expressive Vocalizations (ExVo) challenges *(Baird et al., 2022)*, supports this direction, showing that emotion annotations sampled at 2 Hz (every 500 ms) or derived from brief vocal bursts still

retain meaningful affective content. Building on this evidence, the approach adopted in this work departs from traditional suprasegmental frameworks that rely on handcrafted features and predefined functionals. Instead, it employs an end-to-end neural modeling framework trained directly on raw waveforms, enabling the system to autonomously learn the temporal and spectral abstractions most relevant for emotion inference, without the need for manual feature design or statistical summarization.

This chapter first presents the proposed approach in Section 3.1. In Section 3.2, we demonstrate, using a benchmark corpus of dyadic emotional speech, that end-to-end raw waveform modeling outperforms conventional handcrafted acoustic features for short-segment emotion recognition. Building on these findings, the approach is further extended to model the emotional dimensions of valence and arousal in a stress-inducing free-speech scenario (Section 3.3) and to non-linguistic vocalizations, or vocal bursts (Section 3.4). Section 3.5 introduces analytical methods to examine the nature of emotion-related information captured by the end-to-end framework from short speech segments. Finally, Section 3.6 provides a summary of the chapter.

3.1 Proposed approach for short speech segments modeling

For modeling short speech segments, we employed a raw waveform-based neural architecture previously proposed and evaluated in several paralinguistic and speech-related tasks, including speech recognition *(Palaz et al., 2019)*, speaker verification *(Muckenhirn et al., 2018a)*, gender recognition *(Kabil et al., 2018)*, and depression detection *(Dubagunta et al., 2019)*. In this framework, the raw audio waveform is directly processed by a series of convolutional layers to learn low-level and mid-level representations, which are subsequently fed into a multilayer perceptron (MLP) classifier. We utilized the same network configuration as in the depression detection study *(Dubagunta et al., 2019)*, comprising four convolutional layers followed by a single hidden-layer MLP, along with identical hyperparameter settings. Figure 3.1 illustrates the proposed end-to-end framework adopted for speech emotion recognition task. The input to the neural network is a raw-speech waveform of duration W_{Seq} ($\approx$ 250 ms), which is processed by N convolution layers followed by a MLP to output speech emotion class conditional probabilities. Similar to conventional short-term spectral processing, the speech segment is shifted by 10 ms to estimate class conditional probabilities for the next frame and so on.

During the training phase, the neural network is trained with a frame-level cross entropy error criterion. For inference as demonstrated in Figure 3.1(C), class conditional probabilities estimated for each frame (P_{f_n}) are summed and normalized by the number of frames (i.e mean aggregation) to estimate utterance-level speech-emotion class conditional probabilities. The decision is then taken by selecting the class with maximum probability. In this case, no segment selection is done. The final decision can be seen as soft voting taken on frame level decisions.

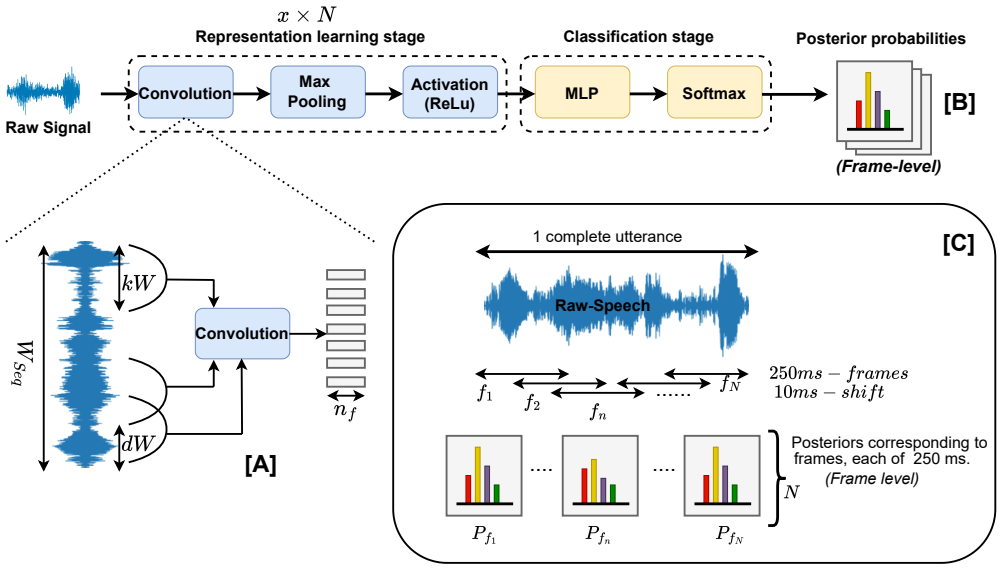


Figure 3.1: Illustration of the proposed speech emotion recognition framework. [A] depicts the processing in the first convolutional layer, where kW denotes the kernel width, dW the kernel shift, and n_f the number of convolutional filters. [B] shows the approach for aggregating frame-level probabilities to perform speech emotion classification. [C] illustrates the segmentation process used to create short segments from an utterance. The input to the neural network is a 250 ms speech segment, and P_{f_n} represents the class-conditional emotion probabilities estimated for each frame.

We evaluate the approach on three tasks: (1) categorical emotion recognition in dyadic conversations; (2) continuous valence/arousal prediction in a stress-inducing, free-speech setting; and (3) continuous emotion prediction for non-linguistic vocalizations (vocal bursts). These tasks jointly assess robustness across speech types and label taxonomies.

3.2 Categorical emotion recognition: dyadic conversations

To model dyadic conversation for the task of SER we used the IEMOCAP American English dataset *(Busso et al., 2008)*, a widely used benchmark corpus in speech emotion research. To be consistent with previous studies *(Rozgić et al., 2012; Xia and Y. Liu, 2015; Ghosh et al., 2016; Neumann and N. T. Vu, 2019a)*, we resorted to the samples from four basic emotion categories- *angry*, *happy*, *neutral* and *sad* with a total of 5531 utterances (with 1103, 1636, 1708 and 1084 utterances each, respectively) by merging the samples from the class *excited* with *happy*. Similar to previous studies on this corpus *(Rozgić et al., 2012; Xia and Y. Liu, 2015; Ghosh et al., 2016; Neumann and N. T. Vu, 2019a)*, we conducted speaker-independent experiments following the leave-one-session-out methodology for training.

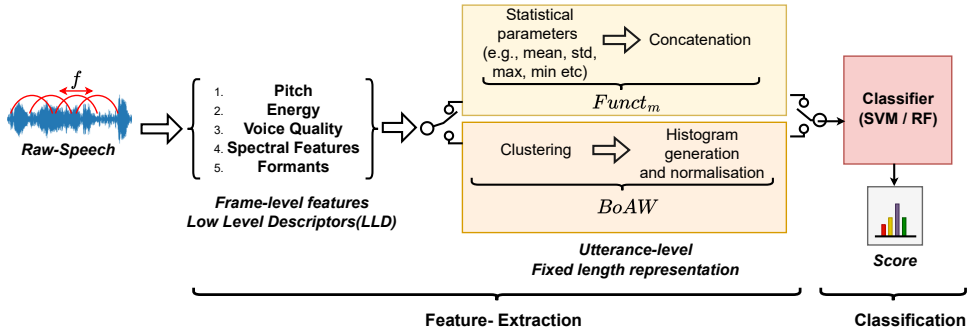


Figure 3.2: Schematic of a conventional handcrafted feature pipeline for paralinguistic analysis. Raw speech is processed over small frames (f), from which Low-Level Descriptors (LLDs) are extracted. These frame-level features are transformed into fixed-length utterance-level representations through either statistical functional concatenation ($Funct_m$) or Bag-of-Audio-Words (BoAW). The resulting vector is processed by a classifier (e.g., SVM or RF) to produce a final prediction score.

3.2.1 Baseline systems

For modeling the acoustic information for emotion classification, we utilize knowledge-based feature representations provided by the openSMILE toolkit *(Eyben et al., 2010)* alongside state-of-the-art acoustic embeddings from WAV2VEC2 *(Baevski et al., 2020)*. The general architectural flow for the handcrafted-based feature pipeline is illustrated in Figure 3.2. Detailed descriptions of these features are provided in Section 2.5.

COMPARE *(Schuller et al., 2013b)* handcrafted frame-level and turn-level feature representations were used in our experimental study. Two configurations of COMPARE features were used in our experiments: COMPARE$_{LLD}$ - 65 + 65 = 130 low–level descriptor (*LLDs*) for frame-level representation and COMPARE$_{LLD\times F}$ - 6373 static turn-level features resulting from the computation of functionals (statistics) over *LLD* contours. We also conducted experiments using EGEMAPS *(Eyben et al., 2016)* 23 dimensional frame-level representations. In order to map frame-level representations into fixed-length turn-level acoustic feature vectors, we use the Bag-of-Audio-Words (BOAW) approach. In our experimental study three configurations of BOAW were used: 500 + 500 = 1000 codebooks for BOAW(COMPARE$_{LLD}$) (500 codebook vectors each for 65 *LLDs* and their delta coefficients) and 1000 codebook vectors for BOAW(EGEMAPS) representation. We also built the BOAW(WAV2VEC2) system based on 768 dimensional wav2vec2.0 Baevski et al., 2020 features obtained from raw speech, using 500 codebook vectors.

Further, the speech emotion classification task was carried out using these fixed-length turn-level acoustic feature vector representations by training support vector machine (SVM) and random forest (RF) classifiers.To establish strong baseline systems, the classifiers built on handcrafted features were optimized through hyperparameter tuning using a grid search approach.

3.2.2 Short-segment based systems

As illustrated in Figure 3.3, we study two approaches: (1) computing the frame-level handcrafted features every 10 ms from the raw audio signal and feeding them as input with temporal context (+12 preceding and following frames) to a MLP to classify emotion at frame level (Figure 3.3.A). (2) feeding raw audio signal of 250 ms every frame to a CNN convolution layer to classify emotion at frame level (Figure 3.3.B). For both, approaches the output frame level probabilities are aggregated at the utterance level to make the final decision. We study these two approaches in comparison with conventional utterance/turn-level speech segment modelling.

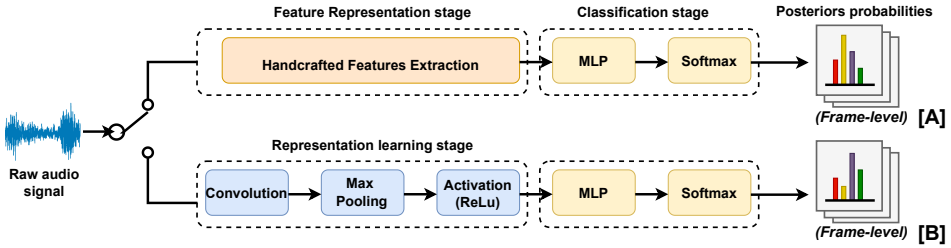


Figure 3.3: Illustration of the proposed approach for modelling short segments of speech. [A] showing the approach of using handcrafted features with a short segment context. [B] showing the approach of directly modelling a short segment of raw-audio signal.

Handcrafted feature-based modelling: For this study we resorted to ComParE$_{LLD}$ and eGeMAPS frame-level feature representations, consisting of feature dimensions 130 and 23 respectively. Using the frame-level features we created short-segments of handcrafted features with a context of 250ms. Each frame in ComParE$_{LLD}$ and eGeMAPS is based on a 10ms analysis window and a context of 12 preceding and succeeding frames for a total of 25 frames. These handcrafted temporal-context based features are used as input to the MLP. The number of layers and hidden nodes was decided based on the cross-validation set.

Raw audio signal modelling: We employed a raw waveform modelling approach previously proposed and studied for speech recognition *(Palaz et al., 2019)*, speaker verification *(Muckenhirn et al., 2018a)*, gender recognition *(Kabil et al., 2018)* and depression detection *(Dubagunta et al., 2019)*, where raw waveform is passed through convolutional layers and then fed to an MLP for classification. We used the same architecture (4 convolutional layers followed by a single hidden layer MLP) and hyper-parameters as used for the depression detection study *(Dubagunta et al., 2019)*. Depending upon the kernel width of the first convolutional layer, we distinguish two CNNs: (a) a kernel width of about 1.8 ms (< 1 pitch period) denoted as Raw SubSeg and (b) a kernel width of about 18 ms (1-5 pitch periods) denoted as Raw Seg.

We split each fold's training data 80:20 into training and cross-validation subsets. Models were optimized with cross-entropy loss using stochastic gradient descent. Whenever validation loss plateaued, we halved the learning rate within the range 10^{-1}–10^{-6}.

3.2.3 Results

Table 3.1 presents the performance of the different systems in terms of unweighted average recall (UAR). Table 3.2 presents different neural network results reported on the same protocol. In Table 3.1, it can be observed that the end-to-end approach yields a UAR of 57.48 and 52.32 for Raw -SubSeg and -Seg systems respectively and outperforms the hand-crafted feature based approach which yields a UAR of 45.88 and 44.36 when modelling short speech segment. The proposed short-segment level modelling end-to-end approach yields performance competitive to conventional utterance-level modeling of speech segments.

Table 3.1: Performance of different systems measured in terms of UAR.

Systems	**Classifier**	**UAR**(↑)
Utterance level modelling		
ComParE$_{LLD\times F}$	SVM	56.57
ComParE$_{LLD\times F}$	RF	58.23
BoAW(ComParE$_{LLD}$)	SVM	56.63
BoAW(ComParE$_{LLD}$)	RF	57.71
BoAW(eGeMAPS)	SVM	55.40
BoAW(eGeMAPS)	RF	55.90
BoAW(Wav2vec2)	SVM	53.7
BoAW(Wav2vec2)	RF	56.0
Short-segment level modelling		
ComParE$_{LLD}$	MLP	45.88
eGeMAPS	MLP	44.36
Raw SubSeg	CNN-MLP	57.48
Raw Seg	CNN-MLP	52.32

It is worth mentioning that the utterance level results for ComParE features and BoAW word representations are comparable to those reported in the literature *(Amiriparian et al., 2021)*. It is interesting to note that the proposed CNN-based raw waveform modelling approach outperforms similar recent approaches that model long speech segments (from Table 3.2).

Table 3.2: Performance of previously reported systems measured in terms of UAR and Weighted Accuracy (WA); Utterance level (UL)

Method (Feature) – Duration	**Metric**	**% (↑)**
Att. CNN (logMel) – 7.5s *(Neumann and T. Vu, 2017)*	WA	56.1
DBN-ivector (MFCC) – UL *(Xia and Y. Liu, 2016)*	WA	57.2
CNN+LSTM (raw aud.) – 6s *(J.-L. Li et al., 2020)*	UAR	52.8
TDNN (MFCC) – 4s *(Kumawat and Routray, 2021)*	UAR	58.6

The hand-crafted feature based systems in Table 3.1 and Table 3.2 together suggest that hand-crafted feature based approaches need long segments for optimal performance. Together,

these results demonstrate that the proposed end-to-end approach is able to effectively model emotion discriminating information from 250 ms of speech.

3.2.4 Neural embeddings based systems

Previous results demonstrate that the neural network is indeed learning information from 250 ms of speech that is indicative of speech emotion, here we investigate an approach where an utterance level representation is obtained from frame level neural embeddings (NN-EMBEDDINGS) either by computing functionals (Funct) such as mean (m) and standard-deviation (sd), or by obtaining a BOAW and classifying the utterance level representation using SVM or RF classifiers.

The results of the evaluated systems are presented in Table 3.3. These findings clearly demonstrate that making decisions at the frame level, as proposed in our approach, yields performance comparable to that obtained using utterance-level representations constructed from frame-level embeddings, a strategy that is commonly adopted in the speech emotion recognition literature.

Table 3.3: Performance of different systems measured in terms of UAR.

Systems	**Classifier**	**UAR**(↑)
Proposed systems - SubSeg ($kW_1 = 30$ samples)		
Raw-CNN	Softmax	57.48
BoAW(NN-EMBEDDINGS)	SVM	56.62
BoAW(NN-EMBEDDINGS)	RF	57.03
Funct_m(NN-EMBEDDINGS)	SVM	56.77
Funct_m(NN-EMBEDDINGS)	RF	56.65
$\text{Funct}_{m,sd}$(NN-EMBEDDINGS)	SVM	56.38
$\text{Funct}_{m,sd}$(NN-EMBEDDINGS)	RF	55.75
Proposed systems - Seg ($kW_1 = 300$ samples)		
Raw-CNN	Softmax	52.32
BoAW(NN-EMBEDDINGS)	SVM	53.02
BoAW(NN-EMBEDDINGS)	RF	53.13
Funct_m(NN-EMBEDDINGS)	SVM	50.67
Funct_m(NN-EMBEDDINGS)	RF	50.06
$\text{Funct}_{m,sd}$(NN-EMBEDDINGS)	SVM	50.90
$\text{Funct}_{m,sd}$(NN-EMBEDDINGS)	RF	51.04

3.3 Continuous emotion recognition: stress-inducing, free-speech scenario

Building upon the findings presented in the previous section, it was observed that embeddings extracted from neural networks trained on short speech segments are effective for classify-

ing categorical emotion labels. To further examine the representational capacity of these embeddings, we extended our investigation to a cross-corpus and cross-language scenario. Specifically, we evaluate whether features learned from short segments are sufficiently expressive to predict the continuous emotional dimensions of valence and arousal. This study was conducted within the experimental framework of the ACM MuSe Challenge *(Amiriparian et al., 2022)*, focusing on the MuSe-Stress task.

The MuSe-Stress task is derived from the multimodal Ulm-Trier Social Stress Test (Ulm-TSST) corpus, which captures participants in a stress-inducing, free-speech context following the standardized Trier Social Stress Test (TSST) protocol *(Kirschbaum et al., 1993)*. In this setup, participants simulate a job interview situation by delivering a five-minute spontaneous oral presentation after a brief preparation period. The dataset comprises approximately six hours of audio recordings from 69 participants (49 female), aged between 18 and 39 years in German language. Each recording has been continuously annotated by three independent raters for the emotional dimensions of valence and arousal at a sampling rate of 2 Hz.

The objective of this task is to predict emotional variations in a time-continuous manner. Following the data protocol established by the challenge organizers, the Ulm-TSST corpus was partitioned into training, development, and test sets containing 41, 14, and 14 subjects, respectively. For model optimization, we further subdivided the training set into two subsets: 32 subjects for training and 9 subjects for validation during hyperparameter tuning. Task performance was evaluated using the Concordance Correlation Coefficient (CCC) for both arousal and valence dimensions.

3.3.1 Methodology

Figure 3.4 illustrates the proposed methodology. In this approach, frame-level neural embeddings are first extracted using pre-trained networks from the input acoustic signal. For each 500 ms segment, a fixed-length representation is then derived by computing statistical functionals specifically, the mean and standard deviation over the embeddings. These representations are subsequently fed into a neural network designed to estimate the emotional dimensions of valence and arousal. Given the sequential nature of the task, the valence-arousal estimation module is implemented using a LSTM-RNN. To ensure a fair comparison, we employ the same LSTM-RNN architecture and training configuration as provided in the baseline system by the challenge organizers, without any architectural modifications. Following the baseline protocol, separate models are trained for valence and arousal prediction, enabling a direct evaluation of our proposed embeddings against the baseline feature representations.

3.3.2 Baseline systems

Two baseline systems, as provided by the challenge organizers, were considered. The first employed the eGeMAPS feature set *(Eyben et al., 2016)*, comprising 88-dimensional hand-

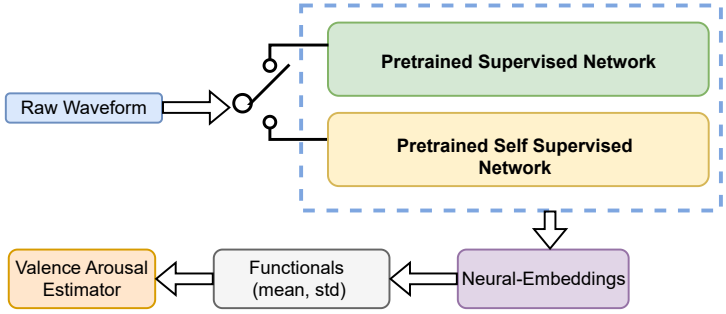


Figure 3.4: Proposed neural embedding modelling approaches for MuSe-Stress task.

crafted acoustic descriptors widely used in paralinguistic studies. Functional features were extracted at 2 Hz using a 5-second analysis window. The second, a DEEPSPECTRUM-based system, utilized the DenseNet121 *(G. Huang et al., 2017)* architecture pre-trained on ImageNet *(Russakovsky et al., 2015)*. Audio signals were converted into Mel-spectrograms with 128 frequency bands and rendered using the viridis colormap. The resulting spectrograms were input to DenseNet121, and activations from the final pooling layer were extracted as 1024-dimensional embeddings using a 1-second window and a 500 ms hop size.

3.3.3 Short segment neural embedding based systems

We used the short-segment based CNN network as presented in Section 3.2 which is trained for classifying four emotion categories namely anger, happy, sad, neutral in an end-to-end fashion by taking 250ms of raw-speech signal as input. This system is referred to as RAW(SER).

3.3.4 Speech Foundation Model based systems

In our experiments, we employed two categories of self-supervised learning (SSL) embeddings: general-purpose audio representations and full-stack speech processing representations. The general-purpose embeddings were obtained from COLA *(Saeed et al., 2021)*, which was trained on the AudioSet *(Gemmeke et al., 2017)* corpus using a contrastive SSL framework. The full-stack speech representations were derived from HUBERT *(W.-N. Hsu et al., 2021)* and WAVLM *(S. Chen et al., 2022)*, both pre-trained on approximately 94,000 hours of speech data.

3.3.5 Results

From Table 3.4, it is evident that the proposed acoustic embedding system, RAW(SER), outperforms the strongest DEEPSPECTRUM baseline on the test set for both valence and arousal. Overall, these embeddings deliver superior results compared to the baseline systems. Furthermore, it is noteworthy that the RAW(SER) embeddings, obtained from a model trained on the English IEMOCAP corpus *(Busso et al., 2008)*, generalize well to the German MuSe-

Stress dataset. This cross-lingual generalization highlights the robustness of our training strategy, which models fine-grained, sub-segmental speech units (around 250 ms), enabling language-independent representations. Remarkably, despite being only 20-dimensional, the RAW(SER) embeddings achieve the highest standalone performance on the combined score. Standalone HUBERT and WAVLM features also perform competitively, with the latter surpassing the baseline DEEPSPECTRUM system. Furthermore, it is noteworthy that combining the 20-dimensional RAW(SER) embeddings with Speech Foundation Model (SFM)-based embeddings enhances the performance of standalone SFM systems. While COLA represents the weakest standalone acoustic embedding, its fusion with RAW(SER) yields a 17.22% relative improvement, demonstrating that these representations are complementary and synergistic for the task at hand.

3.4 Continuous emotion recognition: non-linguistic vocalizations

Building upon the investigation of short-segment modeling in continuous emotion prediction, we further evaluated the proposed approach in the context of non-linguistic vocalizations. This study was conducted within the framework of the ICML ExVo Challenge and Workshop *(Baird et al., 2022)*. The ExVo dataset *(Cowen et al., 2022)* comprises 59,201 recordings of vocal bursts (VB) produced by 1,702 speakers, amounting to approximately 37 hours of data. The recordings were collected from speakers aged between 18 and 39 years across four countries: USA, China, Venezuela, and South Africa.

Each vocal burst was evaluated by a group of listeners who rated the perceived intensity of its emotional expression on a continuous scale ranging from 1 to 100. Every sample is associated with intensity scores (ranging from $[0, 100]$) across ten emotion categories: (1) amusement,

Table 3.4: *CCC* scores (↑) on the development and test sets across different systems. For the development set, results are reported as mean ± std over five random seeds, with the best scores highlighted. Combined results represent the mean of arousal and valence test *CCC*s for each feature set. The symbol "+" indicates early fusion (feature concatenation), and *ndims* refers to the feature dimensionality.

		Arousal [CCC]		**Valence** [CCC]		**Combined** [CCC]
Features	ndims	Development	Test	Development	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
Proposed systems						
RAW(SER)	20	0.3404 (0.2986 ± 0.0311)	0.4338	0.5548 (0.5403 ± 0.0116)	**0.5134**	0.4736
COLA	1280	0.3770 (0.3480 ± 0.0266)	0.4764	0.5572 (0.5268 ± 0.0310)	0.3028	0.3896
HUBERT	2048	0.2622 (0.2388 ± 0.0155)	0.4833	0.5098 (0.4853 ± 0.0161)	0.4309	0.4571
WAVLM	2048	0.2842 (0.2599 ± 0.0183)	0.4462	0.4672 (0.4381 ± 0.0240)	0.4874	0.4668
RAW(SER)+COLA	1300	0.3818 (0.3593 ± 0.0241)	**0.5111**	0.5528 (0.5429 ± 0.0082)	0.4023	0.4567
RAW(SER)+HUBERT	2068	0.3144 (0.3063 ± 0.0063)	0.4724	0.4941 (0.4630 ± 0.0185)	0.4907	**0.4815**
RAW(SER)+WAVLM	2068	0.3114 (0.2924 ± 0.0134)	0.4354	0.4587 (0.4500 ± 0.0065)	0.4648	0.4501

(2) awe, (3) awkwardness, (4) distress, (5) excitement, (6) fear, (7) horror, (8) sadness, (9) surprise, and (10) triumph. These emotions were selected based on their prevalence within the dataset to encompass both broad affective distinctions (e.g., amusement vs. fear) and fine-grained contrasts (e.g., fear vs. horror), consistent with the notion that emotions occupy a high-dimensional, continuous space.

The objective of this task was to jointly predict three aspects from each vocal burst, (i) the average intensities of ten perceived emotions, (ii) the speaker's age, and (iii) the speaker's native country, thereby formulating a multi-task learning problem. The experimental setup followed the official data partitions for training, validation, and testing as defined by the challenge organizers. Model performance was assessed using the competition's task-specific evaluation metrics:

1. *Concordance Correlation Coefficient (CCC)* — computed for each emotion and averaged to evaluate emotion recognition performance.
2. *Mean Absolute Error (MAE)* — used to assess accuracy in age prediction.
3. *Unweighted Average Recall (UAR)* — used for the native country classification task.

An overall multi-task learning score, S_{MTL}, was defined as the harmonic mean of these metrics, computed as:

$$S_{MTL} = \frac{3}{(1/CCC + MAE + 1/UAR)}$$

3.4.1 Methodology

Figure 3.5[A] provides an overview of the proposed approach. In this framework, frame-level neural embeddings are first extracted using pre-trained networks from the acoustic signal. Subsequently, fixed-length representations are derived for each 500 ms segment by computing the mean and standard deviation over the embeddings. These representations are then fed into a multi-task learner block, as illustrated in Figure 3.5[B], which jointly estimates emotional intensity, speaker age, and country. The multi-task learner module (Figure 3.5[B]) is a neural-network block with two hidden layers, with 128 neurons in the first hidden layer followed by 64 neurons in the second layer, with leaky-ReLU activation used for both the layers. The Adam optimization method *(Kingma and Ba, 2015)* is used for training the network. The loss-functions used are: (1) Mean Squared Error (MSE) for age and emotion detection (regression-based tasks) and (2) cross-entropy loss for native-country prediction (classification based task). The mean of these loss functions is calculated for the target. We developed systems with two different configurations: (a) '**Sys-1**' where the network configuration is same as the organizers, and (b) '**Sys-2**' where we doubled the number of neurons in the hidden layers, that is 256 neurons for the first hidden layer and 128 neurons for the second.

To assess the effectiveness of our short-segment emotion modeling approach, we further compare its performance against state-of-the-art (SOTA) self-supervised neural embedding

systems and the baseline models provided by the challenge organizers.

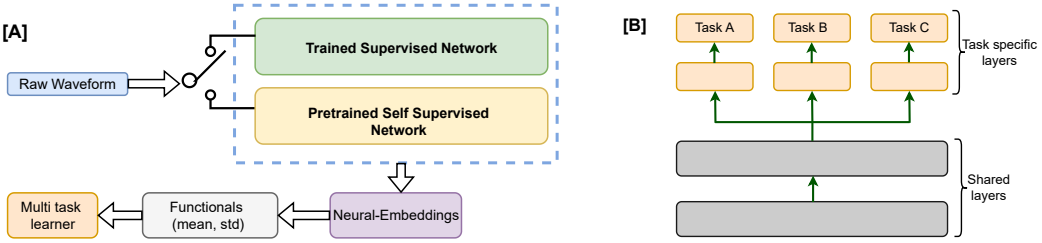


Figure 3.5: [A] Illustration of the proposed neural embedding-based approaches. [B] Overview of the hard parameter-sharing multi-task learner block depicted in [A].

3.4.2 Baseline systems

We considered two baseline systems in our experiments, as provided by the challenge organizers. The first baseline employed the ComParE feature set *(Schuller et al., 2013b)*, consisting of 6373-dimensional handcrafted turn-level acoustic features widely used in paralinguistic studies. The second baseline utilized deep spectral representations extracted using the DeepSpectrum toolkit *(Amiriparian et al., 2017)*, which produces 4096-dimensional embeddings. For this system, the default configuration were used as provided by the DeepSpectrum authors, employing the VGG-19 network *(Simonyan and Zisserman, 2015)* pre-trained on ImageNet *(Russakovsky et al., 2015)*, with spectrograms rendered using the viridis colormap.

3.4.3 Short segment neural embedding based systems

We developed an end-to-end short-segment based CNN network which takes 250ms of raw-waveform as proposed in Section 3.1. The system was trained for a ten-class classification task using the ExVo data, and is referred as Raw(ExVo). We followed a hard-label approach: for each ExVo audio file we picked the categorical emotion based on the highest score provided by the raters to the corresponding emotion category. The network comprises of four convolutional layers followed by a fully connected layer with ten nodes and an output unit consisting of ten emotion classes. The output layer has softmax activation, while all the layers has ReLU activation.

To train the neural network we split the train-set into training and cross-validation subsets in 90:10 ratio. This cross-validation subset was used for hyperparameter tuning. The ExVo provided validation set was used for inference. This training method made sure that the network trained for generating embeddings is trained in a speaker-independent fashion as per the baseline protocols *(Baird et al., 2022)*. The networks were trained using cross-entropy loss with stochastic gradient descent. The learning rate was halved, in the range 10^{-1} to 10^{-6}, between successive epochs whenever the validation-loss stopped reducing.

The frame-level neural embeddings of dimension 10, are extracted before the activation layer

of the fully-connected layer. A fixed-length utterance-level representation is obtained by computing functionals (mean and standard deviation) of the frame-level neural embeddings. This makes these embeddings of dimension 20 (10 mean + 10 std) each.

3.4.4 Speech Foundation Model based systems

To derive SFM embeddings, we make use of the following publicly available SOTA pre-trained SFM systems: WAV2VEC2 *(Baevski et al., 2020)*, HUBERT *(W.-N. Hsu et al., 2021)*, and WAVLM *(S. Chen et al., 2022)*. These systems are among the top three performing networks for the SUPERB challenge *(Yang et al., 2021)*, a SFM benchmark challenge for the speech processing tasks.

In line with the proposed methodology, detailed in Section 3.4.1, utterance-level representations were extracted for all the aforementioned systems and subsequently provided as input to the multi-task learning block (Figure 3.5[B]) to train the system.

3.4.5 Results

Table 3.5 presents the results across all sub-tasks. The RAW(EXVO) system outperforms the baseline COMPARE feature set for both the emotion and age prediction sub-tasks, although its performance on the country classification task remains comparatively weaker.

Table 3.5: Experimental results based on the ExVO data. Reporting scores for the best seed and standard deviation from 5 seeds. Results include mean *CCC* across the 10 (Emo)tion categories, *UAR* for the 4-class (Cou)ntry recognition task (chance level of 0.25 *UAR*), and *MAE* for the age estimation task. S_{MLT} denotes harmonic mean between these metrics. The symbol "+" indicates early fusion (feature concatenation), and *ndims* refers to the feature dimensionality.

Systems	**ndims**	**Config.**	**Emo-CCC**(↑)	**Cou-UAR**(↑)	**Age-MAE**(↓)	S_{MLT}(↑)
ExVo Baseline						
COMPARE	6373	Sys.1	0.416	0.506	4.222	0.349 ± 0.003
DEEP SPECTRUM	4096	Sys.1	0.369	0.456	4.413	0.322 ± 0.003
short-segment based Raw-Wav system (trained on ExVo data)						
RAW (EXVO)	20	Sys-1	0.454	0.331	3.953	0.327 ± 0.006
RAW (EXVO)	20	Sys-2	0.469	0.328	3.805	0.334 ± 0.005
Self-Supervised based representations (Pre-trained networks)						
WAVLM	2048	Sys-1	0.523	0.542	4.094	0.382 ± 0.006
WAVLM	2048	Sys-2	0.548	0.536	4.008	0.390 ± 0.009
HUBERT	2048	Sys-1	0.513	0.508	3.864	0.385 ± 0.004
HUBERT	2048	Sys-2	0.518	0.508	3.782	0.391 ± 0.010
WAV2VEC2	1536	Sys-1	0.390	0.379	3.903	0.330 ± 0.008
WAV2VEC2	1536	Sys-2	0.385	0.376	3.895	0.328 ± 0.005
Early Fusion Experiments						
RAW (EXVO) + WAVLM	2068	Sys-1	0.546	0.542	4.150	0.383 ± 0.006
RAW (EXVO) + WAVLM	2068	Sys-2	0.556	0.522	4.057	0.386 ± 0.004

From Table 3.5, it can be observed that WAVLM and HUBERT achieve the best overall multi-task learning score (S_{MTL}) for a stand-alone system. Specifically, WAVLM demonstrates superior performance on the emotion and country prediction sub-tasks, whereas HUBERT attains the highest accuracy on the age prediction sub-task. The Sys-2 configuration further enhances the performance of both models in their respective strengths. In contrast, WAV2VEC2 performs competitively on the age sub-task but falls short on the emotion and country sub-tasks; moreover, its performance slightly degrades under the Sys-2 configuration.

Given the strong performance of WAVLM on the emotion and country tasks and that of RAW(EXVO) on the age task, we developed an early-fusion system combining these two standalone models. The fusion system yields noticeable performance gains on the emotion and country prediction sub-tasks, although no further improvement is observed for the age prediction sub-task.

Overall, our investigations indicate that SSL-based representations generally provide higher overall scores (S_{MTL}) than task-specific neural representations. However, despite its compact 20-dimensional representation and shallow architecture, the RAW(EXVO) model-built on the principle of short-segment emotion modeling-surpasses the handcrafted baselines on overall scores (S_{MTL}) and performs comparably to the SSL-based systems in modeling age-related paralinguistic information.

3.5 Analysis of short term modelling CNNs for SER

Following the results from the previous sections we observe that short-segment CNNs yield competitive systems for modelling raw audio signals. Therefore, it becomes interesting to analyse what information is being learned from the 250 ms audio signal. To investigate this we conduct a two-tiered analysis: (1) We inspect the cumulative frequency response of the first CNN layer; and (2) We generate relevance signals using gradient-based methods and through relevance signals, we probe what features are getting learned. This analysis was carried out on the setup explained in Section 3.2 using the IEMOCAP corpus.

3.5.1 First CNN layer frequency response analysis

To get insight into the information learned by the CNNs, we analyzed the cumulative frequency response of the first convolutional layer F_{cum} as follows *(Muckenhirn et al., 2018a; Palaz et al., 2019)*:

$$F_{cum} = \frac{1}{n_f} \sum_{k=1}^{n_f} \frac{\mathscr{F}_k}{\|\mathscr{F}_k\|_2}$$

where n_f is the number of filters in the first convolution layer, $\mathscr{F}_k$ is the frequency response of filter $f_k \in \{1, \cdots n_f\}$.

The left sub-plot of Figure 3.6 shows the cumulative frequency response for Raw-SubSeg for all the five folds M-1 to M-5. It can be observed that, irrespective of the fold on which the CNNs are trained, the filters give emphasis to 1000 - 4000 Hz frequency region, similar to CNNs trained to classify phones *(Palaz et al., 2019)*. This suggests that Raw-SubSeg is focusing on emotion related information carried at sub-word unit level.

The right sub-plot of Figure 3.6 illustrates the cumulative frequency response for Raw-Seg for all the five folds M-1 to M-5. It is interesting to again observe that, irrespective of the fold on which the CNN is trained, all the CNNs emphasize similar frequency regions. Compared to Raw-SubSeg, it can be observed that the emphasis shift more towards low frequency region. This observation is consistent with speaker verification and depression detection studies *(Muckenhirn et al., 2018a; Dubagunta et al., 2019)*, where the emphasis is given to modeling voice source related information.

The analysis together with the results obtained shows that SubSeg kernel width helps in better modeling emotion class discrimination. This may be attributed to its ability to model both source and system information well when compared to Seg *(Muckenhirn et al., 2019)*.

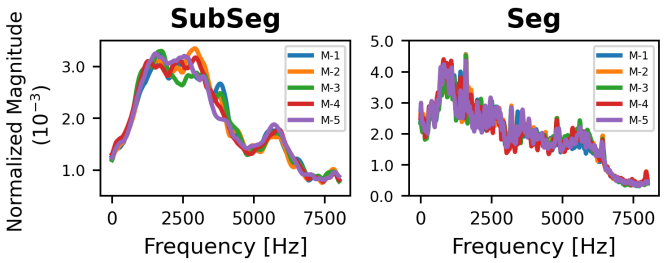


Figure 3.6: Cumulative frequency response of the first convolutional layer for the proposed Raw-CNN models SubSeg (left) and Seg (right). $M-x$ indicates fold x.

3.5.2 Relevance signal analysis

Several recent studies have explored gradient-based techniques for the holistic interpretation of deep feature representations, including guided backpropagation *(Springenberg et al., 2015)*. Extending this line of research, Muckenhirn et al. *(Muckenhirn et al., 2018b; Muckenhirn et al., 2019)* proposed a method to extract temporal and spectral relevance maps, facilitating the visualization and interpretation of task-dependent information learned by CNN-based systems trained directly on raw waveforms. By taking the gradient of the output class with respect to the input signal, relevance signals allow us to measure the impact of perturbations in the input on the output, highlighting crucial discriminative cues in the input. We use relevance signals to gain insights into the information modeled by the proposed methods and show empirical evidence to support our hypothesis.

To ascertain that relevance signals indeed represent crucial discriminative cues in the input

necessary for emotion recognition, we train both the proposed raw CNN models on short-segment relevance signals to classify emotions. Using the trained CNNs, we compute relevance signals for every input in the training data with respect to the ground truth, denoted by SubSeg-Rel and Seg-Rel corresponding to both the proposed SubSeg and Seg CNNs. These relevance signals (SubSeg-Rel and Seg-Rel) were used to train the models. At test time, instead of computing relevance signals only for the ground truth, we generate relevance signals for all four classes and average the predictions of the model on all these four relevance signals. This process is repeated for each fold and UAR is reported, as seen in the bottom two rows of Table 3.6. We can see that each of the models achieves performance quite close to the original SubSeg and Seg models trained on raw audio signals (from Table 3.1). We repeat this procedure for utterance-level modelling by training SVM and RF classifiers on COMPARE$_{LLD}$ features computed from utterance-level relevance signals. From Table 3.6, we can see an improvement in the performance of the SVM classifier when trained using relevance signals obtained from the SubSeg model over the original raw waveform signal (Table 3.1). Together, these results indicate that relevance signals capture input information critical for emotion recognition.

Table 3.6: Performance of different systems on relevance signal, measured in terms of UAR.

Systems	Input Signal	Classifier	UAR
Utterance level modelling			
COMPARE$_{LLD\times F}$	SubSeg-Rel	SVM	57.15
COMPARE$_{LLD\times F}$	SubSeg-Rel	RF	54.06
COMPARE$_{LLD\times F}$	Seg-Rel	SVM	50.57
COMPARE$_{LLD\times F}$	Seg-Rel	RF	54.62
Short-segment level modelling			
Raw-CNN SubSeg	SubSeg-Rel	Softmax	56.37
Raw-CNN Seg	Seg-Rel	Softmax	49.96

To get insights into the information modeled by the proposed CNNs, we rank the top-10 features based on normalized feature importance assigned by RF classifiers trained on COMPARE$_{LLD}$ feature descriptors on the original raw-waveform signal and the relevance signals obtained from the SubSeg and Seg CNNs, as shown in Table 3.7. For the sake of clarity, full feature names were omitted from the table, and the F-Index column indicates the i-th feature from the 0-indexed feature list from the COMPARE header extracted from openSMILE *(Eyben et al., 2010)* toolkit. The “Group" column highlights the broader feature group of the low-level descriptor. The feature grouping has been adopted from *(Schuller et al., 2014)*. In general, the raw-waveform model primarily focuses on spectral low-level feature descriptors (9/10), primarily spectral flux and harmonicity. The SubSeg-Rel based model primarily focuses on cepstral feature descriptors (8/10), more so than the Seg-Rel based model (4/10). It is worth pointing out that the Seg-Rel model puts more emphasis on spectral features (6/10), similar to the raw-waveform model, focusing more on spectral slope descriptors instead of harmonicity. This section contrasts the modeling characteristics of the two CNNs.

Table 3.7: Ranking feature importances from utterance-level RF classifiers trained on COMPARE$_{LLD}$ features obtained from different input signals. F-Index column indicates the i -th feature from the 0-indexed feature list from the COMPARE header extracted from openSMILE. Feature groups as per Schuller et al. (2014)

Raw-waveform		**SubSeg-Rel**		**Seg-Rel**	
F-Index	**Group**	**F-Index**	**Group**	**F-Index**	**Group**
2957	Spectral Flux	5168	Cepstral	5166	Cepstral
1513	Spectral Harmonicity	5166	Cepstral	1523	Cepstral
4138	Spectral (Auditory)	1637	Cepstral	1522	Cepstral
3218	Spectral Harmonicity	1522	Cepstral	2957	Spectral Flux
5496	Spectral (Auditory)	1523	Cepstral	1431	Spectral Slope
6039	Spectral Flux	1587	Cepstral	5097	Spectral Slope
5490	Spectral (Auditory)	5173	Cepstral	5489	Spectral (Auditory)
6035	Spectral Flux	3311	Cepstral	132	Spectral (Auditory)
74	Prosodic	1244	Spectral Flux	5173	Cepstral
6030	Spectral Flux	3712	Voice quality	4144	Spectral (Auditory)

3.6 Summary

This chapter investigated whether emotion-relevant information can be effectively learned from short, sub-segmental segments of speech rather than traditional suprasegmental feature representations. Building on the hypothesis that localized acoustic cues carry affective content, we developed an end-to-end convolutional framework that models raw waveforms of ≈250 ms duration. Experiments across multiple corpora covering dyadic conversations (IEMOCAP), stress-induced free speech (MuSe-Stress), and non-linguistic vocalizations (ExVo), demonstrated that such short-segment models capture salient emotional information and, in several cases, achieve performance comparable to or exceeding utterance-level systems.

For the best-performing raw-waveform-based CNN configuration, namely Subseg, filter analysis revealed that the network predominantly emphasizes the 1000-4000 Hz frequency range, consistent with the behavior of CNNs trained explicitly for phoneme classification *(Palaz et al., 2019)*. In addition, relevance signal analysis indicated a strong emphasis on cepstral feature groups. Taken together, these findings suggest that, when optimizing for emotion recognition, the network tends to exploit phone-relevant information captured from raw waveforms.

This work also establishes that decisions made at the frame level and those derived after temporal aggregation yield equivalent inferences, suggesting that emotional information is locally encoded in short temporal windows.

4 Phonetically aware neural representations for speech emotion recognition

Building on the previous chapter's findings, which showed that short-segment modeling enables networks to capture phone-relevant acoustic cues from raw waveforms, we now focus explicitly on the role of phonetic information in emotion recognition. To this end, this chapter proposes a novel phonetically aware neural modeling framework, aimed at assessing the contribution of phonetically informed representations and revealing how emotional content is expressed through subtle, phone-level variations in speech.

Over the past two decades, different approaches have emerged for speech emotion recognition (SER) *(Schuller, 2018)*, driven by the increasing demand for intelligent systems that can recognize and respond to human emotions, thereby enhancing human-machine interaction. Despite the success of previous frameworks in improving SER performance, most existing work emphasizes turn-level or utterance-level modeling, focusing primarily on acoustic and prosodic attributes aggregated across longer spans of speech *(Schuller et al., 2007; Koolagudi and Rao, 2012; Neumann and T. Vu, 2017; Peng et al., 2021; Pepino et al., 2021; S. Chen et al., 2022)*. However, emotional expression in speech arises from complex and fine-grained interactions among linguistic, phonetic, and articulatory processes *(K. R. Scherer, 2003; K. R. Scherer, 2005; Schuller et al., 2013a)*, motivating efforts to investigate the role of phonetic information in emotion recognition.

Within this evolving landscape, several studies have probed the link between phonetic content and emotional state. C. M. Lee et al. (2004) investigated the impact of emotional coloring on five broad phonetic classes (vowel, glide, nasal, stop, and fricative), they did so by training a phonetic-class based HMM system for each emotional state. The emotion label for the utterance is predicted by first force aligning the input sequence and then comparing the likelihood from each emotion model to determine the emotional state maximizing the likelihood. It is found that vowel sounds as the most effective emotional indicator based on classification performance. Vlasenko and Wendemuth (2013) used a multi-task learning approach, where two sets of HMM-GMM systems were trained to model phonemes based on two emotional states (high and low arousal). For example, the same phoneme /IY/ based upon the emotional state (high or low arousal) is given two different emotion labels. Classification of emotional

state is done during the decoding of the input speech sequence, by taking a majority voting approach for the emotion-labeled phoneme. They demonstrated that phonemes are the smallest possible acoustic units that can classify emotional arousal (high or low). Yenigalla et al. (2018) modeled the phonetic information for SER by mapping the speech signal and/or word transcription into a sequence of phones and then mapping the sequence of phones into an embedding space using Word2vec *(Mikolov et al., 2013)*. Thereby, they explicitly model phonetic information through linguistic knowledge-driven embedding space. Dhamyal et al. (2020) conducted a systematic study to understand the phonetic composition of emotions. Using a self-attention-based emotion classification model they discovered the most 'attended' phonemes for each emotion class. They reported that the distribution of 'attended'-phonemes tend to vary significantly across natural vs acted emotions. A recent study *(Yuan et al., 2021)* followed the approach similar to *(Vlasenko and Wendemuth, 2013; Vlasenko et al., 2014)* but instead used SFM, which enabled training an emotion-dependent model using a small amount of data. The study investigated the best phonetic units for emotion recognition and showed that phonetic units are helpful and should be incorporated in SER. Explicitly modelling phonetic information requires transcripts or a robust phoneme recognition system. However, obtaining speech transcripts incurs overhead, such as the use of human transcribers or the use of speech recognition systems.

These collective efforts provide evidence that phonetic units inherently encode speech emotion information. To advance beyond prior approaches, this chapter introduces an implicit phonetic information modeling framework that captures phonetic information within the learned representation space, without relying on transcriptions.

Section 4.1 presents the study design. Section 4.2 validates the proposed implicit phonetic modeling framework on three benchmark emotional speech corpora. Section 4.3 then extends the analysis to inter-corpus evaluations to assess the generalizability of the proposed approach. It also examines how automatic speech recognition (ASR) accuracy influences the effectiveness of these phonetic embeddings for emotion recognition. Finally, to ensure continuity with the preceding analysis, we evaluate the short-segment phoneme-based model on the same set of tasks presented in the previous chapter. Finally, Section 4.4 concludes the chapter.

4.1 Study Design

4.1.1 Methodology

Figure 4.1 illustrates the approach for implicit modeling of phonetic information in two different manners,

- extracting phonetic embeddings from neural networks specifically trained to classify phones.

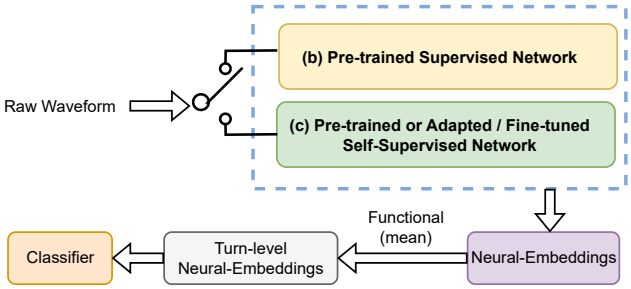


Figure 4.1: Proposed neural embedding-based approaches using system (b) and (c). A detailed explanation of (b) and (c) can be found in Section 4.2.

- extracting phonetic embeddings[I] from SFMs adapted on downstream tasks of phoneme recognition or grapheme recognition.

4.1.2 Datasets and protocols

To validate the proposed approach, we conduct intra-corpus and inter-corpus speech emotion recognition studies. For that purpose, we employ three standard data sets, namely, EMO-DB, IEMOCAP, and MSP-IMPROV.

Berlin Emotional Speech Database : The Berlin Emotional Speech Database (EMO-DB) *(Burkhardt et al., 2005)* covers seven-speaker emotions namely- *anger*, *joy*, *neutral*, *sadness*, *disgust*, *fear*, and *boredom*. The corpus consists of ten (5 male and 5 female) professional actors speaking out ten predefined emotionally neutral sentences. The corpus comprises 900 utterances, where only 493 were marked as 60% natural and 80% assignable by 20 subjects in a listening experiment. We use this subset as suggested in Schuller et al. (2009).

The Interactive Emotion Dyadic Motion Capture : The Interactive Emotion Dyadic Motion Capture (IEMOCAP) *(Busso et al., 2008)* consists of ten (5-male and 5-female) actors over five dyadic sessions performing improvised and scripted scenarios to elicit emotional expressions. In line with previous studies we resort to four emotion classes namely - *angry*, *happy*, *neutral*, and *sad* where we merg the samples from class *excited* with *happy*.

MSP-Improv Database : The MSP-IMPROV *(Busso et al., 2017)* corpus comprises recordings from six spontaneous dyadic sessions enacted by twelve actors (6-male and 6-female) from the University of Texas at Dallas. The database claims to carry more naturalness in the recordings. To adhere to previous works we make use of four emotions namely - *angry*, *happy*, *neutral*, and *sad*.

Table 4.1 summarizes these data sets. To be consistent, for each corpus, we use the protocols that have been used in the literature. More precisely, For EMO-DB, we follow the leave-one-

[I]For the sake of simplicity, we also refer to graphemic/character embeddings as phonetic embeddings, as grapheme and phoneme are related in spoken language.

Speaker-out approach. Whereas, for the other two corpora, we use the leave-one-Session-out methodology. That is, for testing the k-th speaker/session, we train the model on the remaining speakers/sessions. We evaluate the performance of the SER systems in terms of Unweighted average recall (UAR) and Weighted average recall (WAR).

Table 4.1: Data distribution across categorical emotion labels, showing the number of utterances per class.

Database	Content	Ang	Hap	Neu	Sad	Dis	Fea	Bor	**Total**
EMO-DB	German	127	64	78	52	38	55	79	**493**
IEMOCAP	English	1103	1636	1708	1084	-	-	-	**5531**
MSP-IMPROV	English	792	2644	3477	885	-	-	-	**7798**

4.2 Experimental setup and results

4.2.1 System Description

Below, we provide an overview of the systems and their different configurations incorporated for deriving feature representation or neural embeddings by modeling acoustic signals.

(a) Handcrafted feature representation: For the knowledge-based handcrafted feature representation we use COMPARE features *(Schuller et al., 2013b)*. We make use of two configurations of COMPARE features: COMPARE$_{LLD}$ - 65 + 65 = 130 low–level descriptor(*LLDs*) and their delta functions for the frame-level representation and COMPARE$_{LLD\times F}$ - 6373 static turn-level features resulting from the computation of functionals (statistics) over *LLD* contours. Further, we use the Bag-of-Audio-Words (BOAW) approach implemented in the OPENXBOW toolkit *(Schmitt and Schuller, 2017)* to fetch turn-level representations from COMPARE$_{LLD}$ frame-level representation. In BOAW approach 500 + 500 = 1000 codebook vectors were created, 500 for 65 LLDs and 500 for 65 LLDs' delta coefficients. This system is denoted as BOAW(COMPARE$_{LLD}$).

(b) Supervised learning based representation: We utilize an off-the-shelf end-to-end CNN based network for phoneme-classification. The network is trained on the 70hours of AMI (Augmented Multi-party Interaction) Meeting corpus *(Carletta, 2007)* containing 100hours of recordings. The input to the network is a 250ms raw audio signal with a 10ms shift. The network consists of 10 convolutional layers with ReLu activation followed by a fully-connected layer with 1024 neurons, and an output unit with softmax activation for predicting phoneme posteriors. The network is trained for predicting phones based on triphone modeling with cross-entropy loss, stochastic gradient descent and a decaying learning rate. The model provides neural embedding of dimension 1024 corresponding to each 250ms frame. This system is denoted as RAW-CNN(AMI).

(c) Self-supervised learning based representation: We use two different pre-trained self-supervised representation models - Wav2vec2.0 *(Baevski et al., 2020)* and WavLM *(S. Chen et al., 2022)*. The Wav2vec2.0 model is based on a contrastive model approach, the framework

combines contrastive learning with masking. Whereas, WavLM follows a predictive model approach, jointly learning masked speech prediction and denoising in pretraining. For this study, we resort to the base variant of both the models consisting of 12 transformer encoder layers, 768-dimensional hidden states and 8 attention heads comprising 95M and 94.7M parameters for Wav2Vec2.0 and WavLM respectively. Both these models were pre-trained with 960 hours of audio from Librispeech corpus *(Panayotov et al., 2015)*. We utilize these base models to extract the last hidden-state representations with three different settings : (1) The SFMs denoted by WAV2VEC2 and WAVLM. (2) SFMs fine-tuned on TIMIT database *(Garofolo et al., 1993)* for phoneme prediction, denoted by 'SFMs'-FT(TIMIT), and (3) SFMs fine-tuned on 100hours of Librispeech for character classification, denoted by 'SFM'-FT(LIBRI). We adopt S3PRL benchmark framework *(Yang et al., 2021)* for setting (1) and (2), and for setting (3) finetuned models were retrieve from HuggingFace[II].

We use SVM and RF as classifiers. To obtain the best baseline systems, the classifiers for handcrafted features underwent hyperparameter tuning using the grid search methodology. For the neural embedding, the parameters were kept the same, i.e., SVM with a linear kernel and RF with gini criterion, so as to ensure a fair comparison among different embedding spaces.

4.2.2 System Performance

Table 4.2 presents the performance of different systems for the intra-corpus study. Embeddings generated via WAV2VEC2 and WAVLM outperform other systems for all three corpora, with WAVLM delivering the best performance for the SER task. When considering phonetic embeddings, it is interesting to observe that the embeddings derived from RAW-CNN(AMI) network and SFMs adapted for phoneme/character recognition, 'SFM'-FT(TIMIT) and 'SFM'-FT(LIBRI) consistently outperform the knowledge-based handcrafted features on IEMOCAP and MSP-IMPROV, in particular SVM classifier. In the case of German-speaking corpus EMO-DB, however, we do not observe this trend. This suggests that the phonetic embeddings do not generalize well in cross-lingual scenarios. Having said that, when we combine the hand-crafted features and phonetic embeddings through early fusion, we observe that the performance of systems remains steady. In the case of MSP-IMPROV, we observe considerable gains when fusing Group-1 (G-1) and Group-2 (G-2) features, when compared to the standalone G-1 and G-2 feature representations. Finally, we can observe that SVM is able to better model the phonetic embeddings than RF for SER. One possible reason for that could be that RFs partition the feature space using a series of decision trees, which may not be effectively capturing the non-linear relationships between the features in the embedding space.

It is worth mentioning, the performance for handcrafted features reported in Table 4.2 (G-1 features) is comparable to the results previously reported in the literature, for EMO-DB *(Schuller et al., 2009; Eyben et al., 2016)*, for IEMOCAP *(Amiriparian et al., 2021; Purohit et al.,*

[II]Wav2vec2-base-100h: https://huggingface.co/facebook/wav2vec2-base-100h
WavLM-base-100h: https://huggingface.co/patrickvonplaten/wavlm-libri-clean-100h-base

2023b), and MSP-IMPROV *(Busso et al., 2017; Neumann and N. T. Vu, 2019b)*. Similarly, the performance obtained with WAV2VEC2 and WAVLM embedding is consistent with previous studies reported for IEMOCAP and EMO-DB corpus *(Keesing et al., 2021; Pepino et al., 2021; Yang et al., 2021; S. Chen et al., 2022)*.

Table 4.2: Comparison of different feature representations for emotion recognition on three evaluation corpora. Group-1(G-1): Knowledge-based handcrafted features. Group-2(G-2): Supervised learning (SL) based features. Group-3.1(G-3.1): Self-supervised learning (SSL) Wav2vec2 based features. Group-3.2(G-3.2): SSL WavLM based features.

			EVALUATION CORPUS					
			IEMOCAP (4-CLASS)		**MSP-IMPROV (4-CLASS)**		**EMO-DB (7-CLASS)**	
Feature representation	**Dim.**	**Classifier**	**UAR**	**WAR**	**UAR**	**WAR**	**UAR**	**WAR**
Group -1								
COMPARE$_{LLD\times F}$	6373	SVM	58.00	56.51	43.10	55.90	80.20	81.91
COMPARE$_{LLD\times F}$	6373	RF	58.55	57.43	36.21	55.69	66.31	72.97
BoAW(COMPARE$_{LLD}$)	500/500	SVM	57.67	56.62	43.30	55.60	70.85	73.44
BoAW(COMPARE$_{LLD}$)	500/500	RF	58.36	57.41	35.87	55.27	57.19	64.52
Group -2								
RAW-CNN(AMI)	1024	SVM	59.10	58.22	44.33	59.20	77.48	79.72
RAW-CNN(AMI)	1024	RF	52.18	51.89	37.19	56.49	69.16	73.02
Group -3.1								
WAV2VEC2	768	SVM	62.09	61.38	48.60	59.84	83.71	85.40
WAV2VEC2	768	RF	53.27	52.63	38.94	57.19	65.73	72.62
WAV2VEC2-FT(TIMIT)	768	SVM	57.68	56.34	45.89	58.69	67.35	70.79
WAV2VEC2-FT(TIMIT)	768	RF	48.47	48.13	34.25	58.69	49.35	57.81
WAV2VEC2-FT(LIBRI)	768	SVM	62.46	61.25	51.14	61.96	75.24	76.27
WAV2VEC2-FT(LIBRI)	768	RF	51.97	51.00	37.66	53.60	59.03	64.91
Group -3.2								
WAVLM	768	SVM	**64.38**	**63.41**	**54.40**	**64.64**	**87.67**	**88.44**
WAVLM	768	RF	56.99	56.73	38.99	57.94	68.51	74.44
WAVLM-FT(TIMIT)	768	SVM	57.44	56.73	45.69	58.44	63.12	66.33
WAVLM-FT(TIMIT)	768	RF	47.12	46.94	34.26	52.54	44.69	52.33
WAVLM-FT(LIBRI)	768	SVM	60.73	59.61	49.19	61.36	60.22	63.69
WAVLM-FT(LIBRI)	768	RF	48.37	47.89	36.12	52.42	42.59	46.86
Early fusion for selected systems								
G-1+G-2	6373 + 1024	SVM	59.71	58.16	47.63	57.84	80.00	83.40
G-1+G-3.1 (LIBRI)	6373 + 768	SVM	60.62	59.15	50.54	59.72	82.39	84.02
G-1+G-3.2 (LIBRI)	6373 + 768	SVM	60.79	59.29	49.98	59.24	84.25	85.68

4.3 Analysis

4.3.1 Inter corpus training analysis

For the inter-corpus training evaluation, we use Group-1 features for baseline and selected the best-performing phonetic information-based system from each of Group-2 and 3 in Table 4.2. We also conduct early-fusion experiments. We carry out inter-corpus experiments among two databases: IEMOCAP and MSP-IMPROV since they both have the same emotion classes (*angry*, *happy*, *neutral*, and *sad*) and they are based on dyadic conversation. SVM is chosen as the classifier for this analysis.

The relatively lower performance reported in Table 4.3 for inter-corpus training compared to the intra-corpus training performance as in Table 4.2 highlights the challenge associated with generalizing emotion across cross-domain databases. Nevertheless, we can observe that phonetic embedding obtained from the SFMs generalizes across the two corpora better than hand-crafted features. Early fusion of these two features yields a stable performance despite standalone features yielding inferior performance. This indicates that in early fusion the classifier is giving more emphasis to phonetic embeddings than hand-crafted features.

Table 4.3: Performance comparison of different feature representations for the 4-class classification task in the inter-corpus training scheme.

Systems	Dim.	UAR	WAR
Train IEMOCAP Test MSP-IMPROV			
COMPARE$_{LLD\times F}$	6373	38.83	36.86
BoAW(COMPARE$_{LLD}$)	500/500	41.32	39.71
RAW-CNN(AMI)	1024	32.19	52.13
WAV2VEC2-FT(LIBRI)	768	37.69	45.70
WAVLM-FT(LIBRI)	768	44.70	53.53
G1+G2	6373 + 1024	43.31	40.14
G1+G3.1.(LIBRI)	6373 + 768	**46.97**	**48.21**
G1+G3.2.(LIBRI)	6373 + 768	41.37	45.20
Train MSP-IMPROV Test IEMOCAP			
COMPARE$_{LLD\times F}$	6373	41.41	43.33
BoAW(COMPARE$_{LLD}$)	500/500	38.16	40.31
RAW-CNN(AMI)	1024	32.98	35.80
WAV2VEC2-FT(LIBRI)	768	**46.85**	**48.82**
WAVLM-FT(LIBRI)	768	44.66	48.44
G1+G2	6373 + 1024	35.47	38.02
G1+G3.1.(LIBRI)	6373 + 768	45.34	46.71
G1+G3.2.(LIBRI)	6373 + 768	44.58	46.64

4.3.2 Impact of ASR accuracy

From Table 4.2, it can be seen that the SFM-based embedding performance for SER task decreases after we fine-tune the systems with TIMIT for phoneme recognition ('SFM'-FT(TIMIT)). However, it improves (compared to 'SFM'-FT(TIMIT)) when the SFMs are fine-tuned on the 100-h Librispeech corpus ('SFM'-FT(LIBRI)) for character recognition, with the exception of WAVLM for EMO-DB. These initial results suggest that having more data with greater speaker variability may help improve performance for the cross-domain SER task. To further investigate this, we conduct an independent experiment, previously we observed that in the case of

WAV2VEC2-FT(LIBRI) the fine-tuning on the larger corpus (100-h Librispeech) helped attain performance comparable to WAV2VEC2. Therefore, we use the same SFM variant (Wav2vec2.0 base) but fine-tuned on an even larger set, 960-h Librispeech corpus. As expected, this model provides better results for the in-domain task (character recognition) with a lower word error rate of 3.4% on Librispeech 'clean' set in comparison to 6.1% based on 100-h tuning. We then evaluate the extracted embeddings for the SER task. We find that the performance degrades, with UARs of 48.89, 40.18, and 54.54 for IEMOCAP, MSP-IMPROV, and EMO-DB, respectively, when compared to UARs of 62.46, 51.14, and 75.24 (Table 4.2) 100hr fine-tuned net. This indicates that arbitrarily increasing the data does not necessarily yield phonetic embeddings informative for SER. As a by-product, this analysis shows that fine-tuning the SFM model on a large amount of data for speech recognition is making the embedding space less invariant to emotional differences.

4.3.3 Embedding space analysis

While Wav2vec2.0 is originally meant for the ASR task, WavLM is positioned as a full-stack speech processing model *(S. Chen et al., 2022)*. Table 4.2 shows embeddings generated via WAV2VEC2 and WAVLM outperform other systems, this is not surprising, given that these SFM representations are well recognized for carrying rich speech information and have demonstrated strong generalization and competitiveness across various downstream tasks *(Yang et al., 2021)*.

Despite the superior performance of WAVLM based embedding as seen in Table 4.2, it is noteworthy that it exhibits a greater loss of emotional content compared to WAV2VEC2-FT(LIBRI) when fine-tuned for character recognition using 100-h Librispeech. This is evident from the relatively lower performance of WAVLM-FT(LIBRI) compared to WAV2VEC2-FT(LIBRI). This behavior can be attributed to the fact that after fine-tuning, WAVLM emphasize more on spoken content modeling and speaker identity preservation *(S. Chen et al., 2022)*, while discarding the paralinguistic content that carries emotional information. Consequently, the embedding space becomes less expressive of emotions, which can explain the drop in performance for the cross-domain SER tasks.

One can observe from Table 4.2 that the embedding produced by RAW-CNN(AMI) outperforms SFM-based phonetic embedding in the case of EMO-DB. This observation could be explained by the fact that RAW-CNN(AMI) is trained to predict context-dependent tri-phones, which appears to make the embedding space more robust.

Finally, to visualize the embedding space, we compare the representations generated by the SFM before and after fine-tuning on the phoneme recognition task, as illustrated in Figure 4.2. We generate embedding for EMO-DB, since it is a phonetically balanced database. For the case of WAVLM, we observe clusters corresponding to sentence ID's whereas this is not seen for the case of WAV2VEC2. Once the models are finetuned for phoneme recognition these clusters can be seen for both the cases. This suggests that WAVLM is modeling spoken content

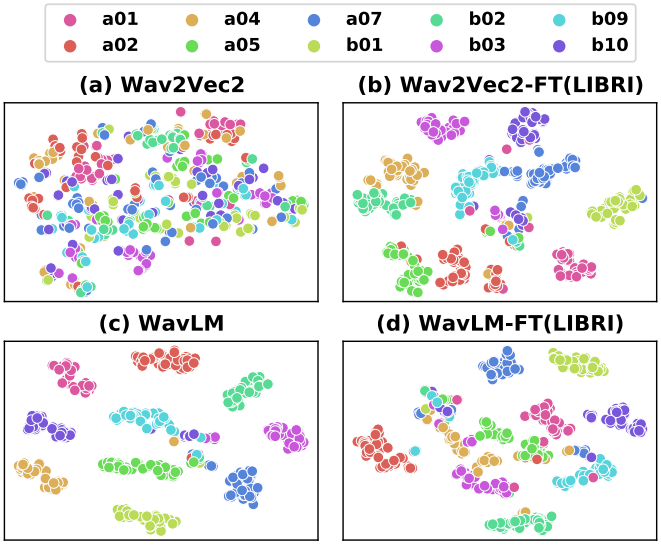


Figure 4.2: t-SNE plots for different embeddings spaces before and after finetuning for the phoneme recognition task. Labels aXX and bXX correspond to text IDs' used in the EMO-DB database.

and speaker identity *(S. Chen et al., 2022)* without any fine-tuning, unlike WAV2VEC2 which may explain WAVLM performance in Table 4.2.

4.3.4 Analysis of the short-segment phoneme-based model

To ensure completeness and continuity with the preceding analysis, we evaluate the supervised short-segment phoneme classification model (RAW-CNN(AMI)), on the same speech emotion corpus discussed in Chapter 3. This comparison allows us to systematically examine the usefulness of short-segment phonetic information modeling in speech emotion recognition across multiple tasks.

Dyadic conversations: The corpus configuration for IEMOCAP and the overall experimental setup follow the same protocol described in Chapter 3, Section 3.2. Using this setup, we evaluate the RAW-CNN(AMI) model. The corresponding utterance-level baseline systems are detailed in Section 3.2.2, while the short-segment-level SER neural architecture is described in Section 3.2.4. For comparison, we employ the superior SUBSEG configuration (see Section 3.1) and aggregate the segment-level embeddings using mean and standard deviation statistics, referred to as RAW(SER). The evaluation results, summarized in Table 4.4, reveal that the RAW-CNN(AMI) embeddings—when modeled using an SVM classifier surpass all baseline systems as well as the short-segment SER model trained directly on IEMOCAP. Notably, despite being trained on the AMI corpus, the RAW-CNN(AMI) system generalizes effectively to emotional speech from a different domain, indicating the robustness and transferability of phoneme-level modeling for SER.

Table 4.4: Experimental results on IEMOCAP corpus. Performance of different systems measured in terms of UAR.

Systems	Classifier	UAR(↑)
Utterance level modelling		
COMPARE$_{LLD\times F}$	SVM	56.57
COMPARE$_{LLD\times F}$	RF	58.23
BoAW(COMPARE$_{LLD}$)	SVM	56.63
BoAW(COMPARE$_{LLD}$)	RF	57.71
BoAW(EGEMAPS)	SVM	55.40
BoAW(EGEMAPS)	RF	55.90
BoAW(WAV2VEC2)	SVM	53.7
BoAW(WAV2VEC2)	RF	56.0
Short-segment level modelling		
RAW(SER)	SVM	56.38
RAW(SER)	RF	55.75
RAW-CNN(AMI)	SVM	59.10
RAW-CNN(AMI)	RF	52.18

Stress-inducing, free-speech scenario: Section 3.3 describes in detail the task definition, the systems employed, the corpus used, and the overall experimental methodology. Here, we assess the effectiveness of the RAW-CNN(AMI) representation for the task on hand. From the results presented in Table 4.5, the RAW-CNN(AMI) embeddings exhibit strong performance on the MUSE–STRESS dataset (German language), despite being derived from a network trained exclusively on an English corpus. This suggests that the learned representations capture language-independent acoustic-phonetic patterns relevant to emotional expression. Furthermore, the results confirm that phonetic-level information is beneficial and complements the emotion recognition task, consistent with the observations reported in previous studies discussed in this chapter.

Table 4.5: Experimental results on MuSe-stress corpus. *CCC* scores (↑) on the development and test sets across different systems. For the development set, results are reported as mean ± std over five random seeds, with the best scores highlighted. Combined results represent the mean of arousal and valence test *CCC*s for each feature set. The symbol "+" indicates early fusion (feature concatenation), and *ndims* refers to the feature dimensionality.

		Arousal [CCC]		**Valence** [CCC]		**Combined** [CCC]
Features	ndims	Development	Test	Development	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
Proposed systems						
RAW(SER)	20	0.3404 (0.2986 ± 0.0311)	0.4338	0.5548 (0.5403 ± 0.0116)	**0.5134**	0.4736
RAW-CNN(AMI)	2048	0.3515 (0.3371 ± 0.0102)	**0.4909**	0.4122 (0.3894 ± 0.0217)	0.4767	0.4838
RAW(SER)+RAW-CNN(AMI)	2068	0.3742 (0.3540 ± 0.0176)	0.4850	0.4081 (0.3804 ± 0.0214)	0.4966	**0.4908**

Notably, among the standalone embeddings, RAW-CNN(AMI) achieves the highest perfor-

mance for emotional arousal prediction. Additionally, the complementary nature of RAW-CNN(AMI) and RAW(SER) is evident from their early-fusion results, which yield a superior overall score, underscoring the advantage of integrating phonetic-aware and emotion-specific representations.

Non-linguistic vocalizations (vocal bursts): The details of the task, experimental setup, system configurations (Sys-1, Sys-2) and methodology are provided in Section 3.4. In this section, we evaluate the efficacy of the RAW-CNN(AMI) representation for classifying emotion intensity in vocal bursts. The corresponding baseline and short-segment-based RAW(EXVO) systems are described in Section 3.4.2 and Section 3.4.3, respectively. The results of this analysis are summarized in Table 4.6. The RAW-CNN(AMI) embeddings achieve performance comparable to the baseline system and surpass the DeepSpectrum results. Furthermore, the combination of the short-segment based systems - RAW(EXVO) and RAW-CNN(AMI) embeddings demonstrates their complementary nature, as evident from the fusion results in Table 4.6. When fused, these embeddings yield a superior overall score (S_{MLT}) compared to the baseline, and the Sys-2 configuration further enhances performance.

Table 4.6: Experimental results based on the ExVO data. Reporting scores for the best seed and standard deviation from 5 seeds. Results include mean *CCC* across the 10 (Emo)tion categories, *UAR* for the 4-class (Cou)ntry recognition task (chance level of 0.25 *UAR*), and *MAE* for the age estimation task. S_{MLT} denotes harmonic mean between these metrics. The symbol "+" indicates early fusion (feature concatenation), and *ndims* refers to the feature dimensionality.

Systems	**ndims**	**Config.**	**Emo-CCC**(↑)	**Cou-UAR**(↑)	**Age-MAE**(↓)	S_{MLT}(↑)
ExVo Baseline						
COMPARE	6373	Sys.1	0.416	**0.506**	4.222	0.349 ± 0.003
DEEPSPECTRUM	4096	Sys.1	0.369	0.456	4.413	0.322 ± 0.003
short-segment based Raw-Wav system (trained on ExVo data)						
RAW (EXVO)	20	Sys-1	0.454	0.331	3.953	0.327 ± 0.006
RAW (EXVO)	20	Sys-2	**0.469**	0.328	**3.805**	0.334 ± 0.005
short-segment based Raw-Wav system (trained on AMI corpus)						
RAW-CNN(AMI)	2048	Sys-1	0.370	0.477	4.210	0.333 ± 0.002
RAW-CNN(AMI)	2048	Sys-2	0.392	0.465	4.179	0.338 ± 0.005
Early Fusion Experiments						
RAW(EXVO) + RAW-CNN(AMI)	2068	Sys-1	0.440	0.480	4.159	0.352 ± 0.003
RAW(EXVO) + RAW-CNN(AMI)	2068	Sys-2	0.452	0.488	4.144	**0.357 ± 0.003**

4.4 Summary

This chapter presented an implicit phonetic modeling framework for speech emotion recognition, wherein phonetic information is captured within learned neural representations rather than through explicit transcriptions. The approach leverages both supervised networks trained for phoneme recognition and self-supervised speech foundation models fine-tuned for phoneme or grapheme classification to extract phonetic embeddings, which are then

modeled at the turn level and classified using SVM or Random Forest backends.

Experimental evaluations on three benchmark emotional speech corpora showed that the proposed phonetic embeddings achieve competitive, and often superior, performance compared to traditional handcrafted acoustic features. However, we observed limited or no additional gains when comparing these embeddings to pre-trained speech foundation models. One possible reason is the recent finding that pre-trained SFMs already tend to encode substantial phonetic information *(Choi et al., 2024)*.

Furthermore, the fusion of phonetic and handcrafted features was shown to enhance robustness across corpora, confirming their complementary nature.The analysis of the phonetic embeddings revealed an inverse relationship between automatic speech recognition accuracy and the emotional discriminability of the derived phonetic embeddings, suggesting that optimizing for linguistic precision may come at the cost of affective sensitivity.

Finally, extending the short-segment phoneme-based model to the tasks introduced in the previous chapter reaffirmed that phonetic-level information provides meaningful and language-independent cues for emotion recognition. Collectively, these findings establish implicit phonetic modeling as an effective, transcription-free pathway for capturing affective information in speech, while opening new directions toward understanding how phonetic granularity influences emotional representation learning.

5 Probing speech foundation models for emotion information recovery

In the previous chapter (Chapter 4), we identified a critical trade-off: as a pretrained Speech Foundation Model (SFM) is fine-tuned for Automatic Speech Recognition (ASR), its capacity to encode paralinguistic information relevant for Speeech Emotion Recognition (SER) diminishes. The reduction in Word Error Rate (WER) achieved through larger-scale ASR fine-tuning corresponds to a gradual loss of affective and speaker-state cues in the learned representations, a trend consistent also reported in prior findings by *(Pepino et al., 2021; Y. Li et al., 2023)*. This observation implies that increasing task specificity enhances linguistic precision at the expense of paralinguistic generality. Building on this insight, the present chapter investigates whether such lost paralinguistic information can be recovered partially or fully through further adaptation of the foundation models.

When employing Foundation Models (FMs) *(Bommasani et al., 2021)* for the downstream tasks, there are two prevalent approaches: (a) full fine-tuning, involving the updating/tuning of all model parameters, and (b) linear probing, where the entire network is frozen, and only the last linear layer (known as the 'head') is tuned for the target task. In the Independent and Identically Distributed (IID) setting, it is known that fine-tuning outperforms linear probing *(Kornblith et al., 2019; Zhai et al., 2019)*. Therefore, usually fine-tuning becomes the de facto approach to yield state-of-the-art performance. Despite the prevalent usage of fine-tuning for adapting a FM, a comprehensive understanding of its underlying mechanisms remains an open question and actively being investigated by the machine learning community *(Kumar et al., 2022; Mukhoti et al., 2023; Zheng et al., 2023)*.

It is well-known that when a model undergoes self-supervised pretraining, employing either contrastive loss *(Chopra et al., 2005)* or reconstruction loss, it learns a robust representation of the input modality (e.g., speech). This quality makes pretrained models a valuable choice as a 'universal' feature extractor *(Yang et al., 2021; A. Mohamed et al., 2022)*. However, after adapting to a specific task, these networks often acquire specialization for that adapted task. It was observed that the fine-tuning process has the potential to distort the patterns learned by the pretrained model on a large corpus, resulting in a decline in the quality of the model's generated outputs *(McCloskey and Cohen, 1989; French, 1999; K. Lee et al., 2019; Zheng et al.,*

2023). We question, does performing a single round of fine-tuning on a pre-trained network distort its representations to the extent that it becomes unsuitable for further fine-tuning for tasks in different domains?

Researchers in the speech community have utilized different SFMs *(Baevski et al., 2020; W.-N. Hsu et al., 2021; S. Chen et al., 2022)* and investigated how representations evolve across layers *(Pasad et al., 2021; Pepino et al., 2021; W. Wu et al., 2023; Pasad et al., 2024)*. Whereas this work aims to investigate the information recovery potential of the FMs. Information recovery, in this context, pertains to the network's capability to attain a comparable state of information encoding. We evaluate this comparability when the pretrained network is directly adapted for a target task or if the pretrained network is initially adapted to an auxiliary task before undergoing further adaptation for the target task, as illustrated in Figure 5.1.

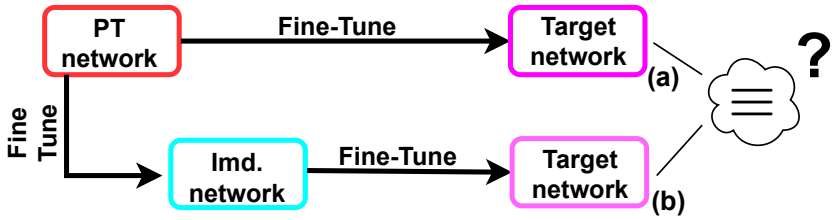


Figure 5.1: Diagram illustrates two pathways to attain the target network, with the same target task. We ask if these target networks (a & b) are equivalent. 'PT' refers to the pre-trained SSL network, and 'Imd.' denotes the intermediary network.

The remainder of this chapter is organized as follows. Section 5.1 outlines the study design and the proposed methodology for probing information recovery in SFMs. Section 5.2 describes the datasets and experimental protocols used in this study. Section 5.3 details the system configurations and presents the main experimental results. Section 5.4 provides a three-tier in-depth analysis of the representational behavior and recovery patterns observed across the model. Finally, Section 5.5 concludes the chapter with key takeaways.

5.1 Methodology and Study design

Figure 5.2 illustrates our methodology for examining information recovery in the FMs. In this study, we propose to systematically model and analyze the representations derived from three specific systems:

1. Modeling the 'universal' embeddings derived from the pre-trained FM for the target task.
2. Fine-tuning the pre-trained FM to an intermediary task-specific system, and subsequently extracting and utilizing these representations for the target task.
3. Further adapting/fine-tuning the intermediary task-specific network acquired in the previous step to the target task and utilizing the corresponding representations.

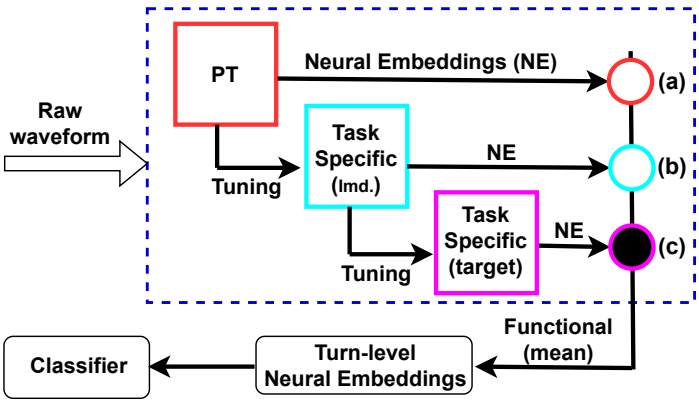


Figure 5.2: Proposed systems: (a) generates pre-trained embeddings, (b) generates intermediary (Imd.) task-specific representation, and (c) generates target task representation. The circles (○) indicate the switching system, while the filled circle (●) denotes the activated switch, directed towards the classifier block.

In this case study we examine our hypothesis, by considering ASR as our intermediary task and ER as our target task and probe the following questions:

1. Is the loss of paralinguistic information a permanent effect, or is there a possibility of its recovery through additional fine-tuning of the ASR network for the ER task? Furthermore, does this result in the same encoding of paralinguistic information as we would observe if we had fine-tuned the pretrained network directly for the target task?

2. If recovery is possible, what are the implications on the decision-making capabilities of the system compared to the network directly finetuned for the target task?

5.2 Dataset and protocols

The Interactive Emotion Dyadic Motion Capture: The Interactive Emotion Dyadic Motion Capture (IEMOCAP) *(Busso et al., 2008)* features ten actors (5 male and 5 female) participating in five dyadic sessions. These sessions involve both improvised and scripted scenarios designed to evoke emotional expressions. For consistency with previous studies, we categorize emotions into four classes: *angry*, *happy*, *neutral*, and *sad*. Notably, we merge samples from the *excited* class with the *happy* class, encompassing a total of 5531 utterances.

MSP-Improv Database: The MSP-IMPROV *(Busso et al., 2017)* corpus comprises recordings from six spontaneous dyadic sessions involving twelve actors (6 male and 6 female) affiliated with the University of Texas at Dallas. The database is known for its emphasis on naturalness in the recorded interactions. In line with prior studies, we adopt four emotion categories: *angry*, *happy*, *neutral*, and *sad* comprising of a total 7798 utterances.

To adhere to previous works, for each corpus, we use the protocols that have been used in the literature. More precisely, we use the leave-one-session-out methodology. That is, for testing the 'k'-th session, we trained the model on the remaining sessions. We evaluate the performance of the SER systems in terms of Unweighted average recall (UAR) and Weighted average recall (WAR).

5.3 Systems and results

5.3.1 System description

(a) Handcrafted feature representation: We employ ComParE features *(Schuller et al., 2013b)*. Two configurations of ComParE features are utilized: ComParE$_{LLD}$, comprising 130 low-level descriptors (*LLDs*) and their delta functions for frame-level representation, and ComParE$_{LLD\times F}$, consisting of 6373 static turn-level features derived from computing functionals (statistics) over *LLD* contours. Additionally, we apply the Bag-of-Audio-Words (BoAW) approach implemented in the openXBOW toolkit *(Schmitt and Schuller, 2017)* to extract turn-level representations from the ComParE$_{LLD}$ frame-level representation. In the BoAW approach, 1000 codebook vectors were created, with 500 for the 65 LLDs and 500 for the delta coefficients of 65 LLDs. This system is denoted as BoAW(ComParE$_{LLD}$).

(b) SFM based representation: We leverage Wav2vec2.0 *(Baevski et al., 2020)* representations. The Wav2vec2.0 model adopts a contrastive learning approach, combining it with masking techniques. In this study, we employ the base variant of the model, which includes 12 transformer encoder layers, 768-dimensional hidden states, and 8 attention heads, totaling 95 million parameters. The model underwent pre-training using 960 hours of audio data from the Librispeech corpus *(Panayotov et al., 2015)*. For this study, we investigate four variations of Wav2vec2.0 :

1. The default pre-trained network, identified as PT.
2. PT fine-tuned specifically for our target task of emotion recognition, denoted by SER.
3. PT fine-tuned for the intermediate ASR task with three configurations, each based on the amount of data used for the fine-tuning process: (a) Fine-tuned with 10 minutes of LibriSpeech data (ASR10), (b) Fine-tuned with 100 hours of LibriSpeech data (ASR100), and (c) Fine-tuned with 960 hours of LibriSpeech data (ASR960).
4. Networks derived from the intermediate ASR task further adapted for our target task (ER), labeled as ASR(x)→SER, where x represents the different ASR models as mentioned above.

ASR-based fine-tuned model checkpoints were retrieved from HuggingFace, ASR10 *(Huggingface, 2023a)* with a WER of 57.81%, ASR100 *(Huggingface, 2023b)* with a WER of 6.1%, and

ASR960 *(Huggingface, 2023c)* with a WER of 3.4% on the clean set of Librispeech. To fine-tune Wav2vec2.0 for Speech Emotion Recognition (SER), we follow the default S3PRL *(Yang et al., 2021)* configuration with minor adjustments. The learning rate is set to 1.0×10^{-5}, using the cross-entropy loss function, the batch size is 4, gradient accumulation is configured at 8, and a random seed value of 1337 is utilized. During Wav2vec2 fine-tuning, the convolution-based encoder blocks are kept frozen, while all the 12 transformer encoder blocks are fine-tuned.

Support Vector Machine (SVM) was utilized as a classifier. We performed hyperparameter tuning for the classifiers associated with handcrafted features using the grid search. For neural embeddings we maintain a consistent linear kernel for SVM, but optimize the values of C and γ parameters through grid search, employing a 5-fold cross-validation split. This approach ensures a fair comparison across various embedding spaces.

5.3.2 System performance

From Table 5.1 it is evident that the SER network outperforms other systems across both databases, as anticipated due to its specific optimization for the ER task. There is a significant enhancement in ER task performance when transitioning from ASR10 to ASR100, characterized by a lower WER for ASR100. However, a subsequent decline in ER performance is observed with ASR960, even with it being a superior ASR system with the lowest WER. This observation aligns with findings reported in prior literature *(Purohit et al., 2023a)*. Upon

Table 5.1: Comparison of different feature representations for emotion recognition on two evaluation corpora.

		Evaluation corpus			
		IEMOCAP (4-class)		MSP-IMPROV (4-class)	
Feature representation	**Dim.**	**UAR ↑**	**WAR ↑**	**UAR ↑**	**WAR ↑**
		G-1: Baseline Features			
ComParE$_{LLD\times F}$	6373	58.00	56.51	43.10	55.90
BoAW(ComParE$_{LLD}$)	500/500	57.67	56.62	43.30	55.60
		G-2: Pretrained network embeddings			
PT	768	56.76	56.26	47.01	58.49
		G-3: Task specific fine-tuning network embeddings			
SER	768	**64.98**	**63.89**	**56.54**	63.41
ASR10	768	55.18	53.59	41.42	58.10
ASR100	768	60.03	58.09	49.25	60.44
ASR960	768	49.38	49.34	36.51	62.22
		G-4: 2 step fine-tuning network embeddings			
ASR10→SER	768	60.93	59.52	52.80	59.91
ASR100→SER	768	64.59	63.68	56.29	**63.99**
ASR960→SER	768	63.57	62.56	55.54	62.72

comparing the performance of PT and ASR100, it becomes evident that the incorporation of phonetic information in the ASR network is, to some extent, more beneficial than the vanilla pre-trained network for the ER task. But, as the ASR network is further optimized for improved ASR performance, there is a notable decline in SER performance. This reaffirms that as the model becomes more optimized for ASR, it tends to lose paralinguistic information which might be undesirable for ASR. Examining the G-4 section of Table 5.1, all the three ASR systems, when further fine-tuned for the SER task, achieve comparable performance to SER in the G-3 section of the table. We do not observe any performance gain through using ASR-based initialization in a two-step fine-tuning process. Nevertheless, the network proficiently regains emotion information, with ASR100→SER demonstrating the most effective recovery, while ASR10→SER exhibits the least recovery. It is worth mentioning that both direct fine-tuning and two-step fine-tuning outperform handcrafted features in terms of performance. Additionally, our results align with previously reported figures in the literature for both IEMOCAP *(Pepino et al., 2021; Y. Li et al., 2023; Purohit et al., 2023b)* and MSP *(Busso et al., 2017; Neumann and N. T. Vu, 2019b; Purohit et al., 2023a)*. It is important to emphasize that our primary focus is not on competing for state-of-the-art results. In this study, we compare and analyze various systems using a similar parameter setup to explore the information recovery potential of FMs.

5.4 Analysis

5.4.1 Effects of two-step fine-tuning on decision outcomes

To further investigate the information recovery within the two-step fine-tuned systems (ASR(x) → SER), we examined the decision outcomes of these systems. We computed the decision mismatches between the predicted labels of the SER network and the ASR(x) → SER networks. The mismatch values are presented in Table 5.2. At first glance, it is noticeable that there is a mismatch of more than 25% for all the networks. It is interesting to point out, in spite of the seemingly close UAR values between ASR100 → SER (64.59%) and SER (64.98%) in Table 5.1, there exists a more than 25% discrepancy in their decision outcomes. Consequently, we do observe a discernible shift in the decision properties.

Table 5.2: Comparison of decision mismatch between the predictions of SER and ASR(X) → SER network.

	IEMOCAP	MSP-IMPROV
Feature representation	**Mismatch %**	**Mismatch %**
ASR10→SER	35.18	34.03
ASR100→SER	27.57	25.10
ASR960→SER	29.66	28.89

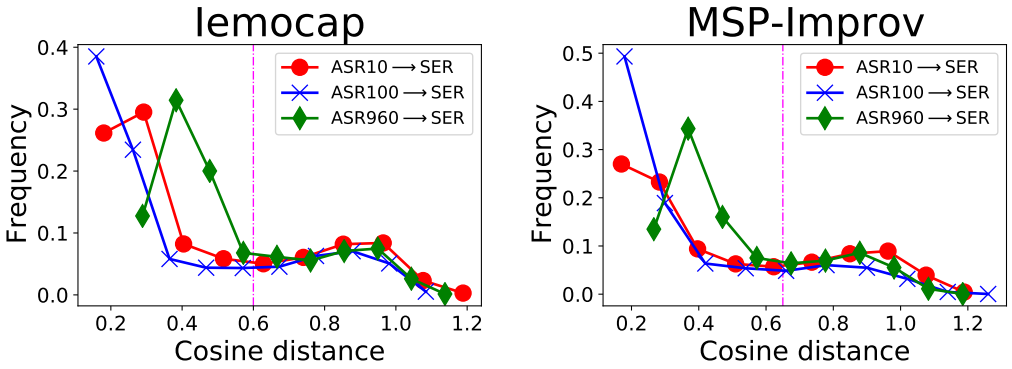


Figure 5.3: Distribution of cosine distance values computed for the last layer representation between SER and ASR(x)→SER network, for all the data points in the corpus. Vertical dashed magenta line indicates the threshold value.

5.4.2 Latent space analysis

In addition to examining the decision outcomes, we analyzed the embedding space by comparing the last layer representations of SER and ASR(x)→SER systems. For this analysis we resorted to the cosine distance formulation, which helps measure the similarity between two non-zero vectors. We calculated the cosine distance between the embeddings generated from SER and ASR(x)→SER systems for all data points in both the corpora used in the study. Figure 5.3 showcases the distribution of cosine distances for all data points using a line-joined histogram plot. The markers (e.g. -•-) on the curves in the subplots of Figure 5.3 represent the center of each of the 10 bins for their respective systems. The distribution of cosine distances between SER and ASR100→SER (denoted by -×- curve) reveals that the majority of data points exhibit lower distances, indicating a higher degree of similarity. In contrast, the diamond head curve (-♦-) representing cosine distance distribution between SER and ASR960→SER shows lower similarity. These observations align with the results presented in Table 5.1. To better assess the alignment of cosine distance with the decisions, we compute decision matches between the predictions of SER and ASR(X)→SER for data points falling within a defined distance threshold. This threshold is represented by a dashed magenta line in the subplots of Figure 5.3, with values set at 0.6 for Iemocap and 0.65 for Msp-Improv. We choose the threshold to encompass more than 70% of the total data points, while maintaining a small cosine distance for analysis. Table 5.3 provides details on the percentage of data points falling within the cosine distance threshold for different systems, along with the decision matches between the SER prediction and ASR(X)→SER prediction for those encompassing data points. The values in the match percentage column of Table 5.3 suggest that the networks can effectively recover the information to a considerable extent.

It is important to highlight that there were a few data points where the cosine distance exceeded one, indicating complete dissimilarity. Upon examining the decision matches, no matches were found for these points. Instances where the cosine distance exceeded one were primarily associated with ASR10→SER (-•- curve), constituting 3.6% of the total data

Table 5.3: % of data falling within the cosine distance threshold, along with the corresponding decision match %.

	IEMOCAP		MSP-IMPROV	
Feature representation	**Data %**	**Match %**	**Data %**	**Match %**
ASR10→SER	70.85	89.00	70.00	85.63
ASR100→SER	75.55	90.69	81.98	88.75
ASR960→SER	70.01	94.07	73.04	89.69

from Iemocap and 5.5% from Msp-Improv. Conversely, the least occurrences were noted with ASR100→SER (-×- curve), accounting for 1.5% in Iemocap and 2% in Msp-Improv.

5.4.3 Attention head analysis

We extract self-attention weights for each head in every layer of the Wav2vec2.0 model for a particular input audio. This results in a 2D float-type array of shape $N \times N$, where N is the frame-wise sequence length of the input audio. The resulting 2D representation, illustrated in Figure 5.4, is referred to as the self-attention map (SAM). Each row in the SAM is a probability distribution representing the attention logits for a specific element to all other elements in the sequence. For our analysis of SAMs, we make use of Equation 5.1, which computes the Bhattacharyya distance (BC) *(Bhattacharyya, 1943)*, providing a symmetric distance measure

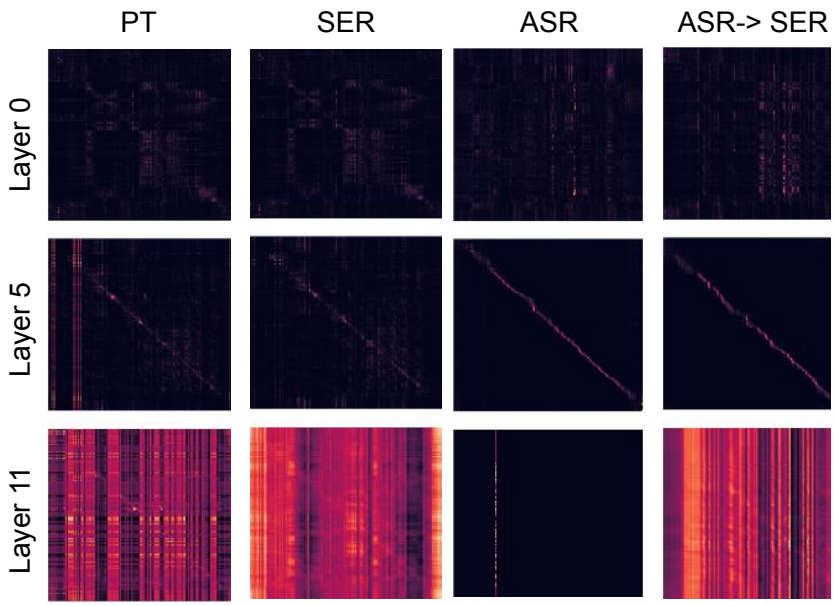
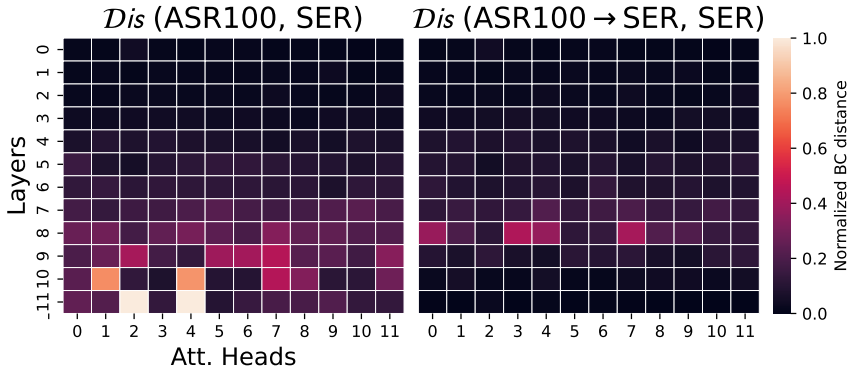
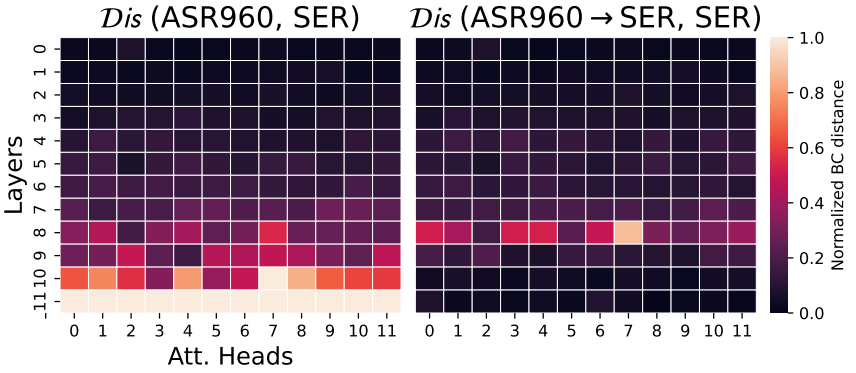


Figure 5.4: Illustration of Self-Attention Maps (SAMs) across different systems, highlighting varied attention head patterns across different transformer layers for the 6th attention head.



(a) Comparing: SER with ASR100 & ASR100 → SER



(b) Comparing: SER with ASR960 & ASR960 → SER

Figure 5.5: Bhattacharyya distance comparison (using Equation 5.1) for the attention heads at different layers for different systems.

between the rows of the SAMs generated by different systems to be compared. Each unit/cell in Figure 5.5 subplots represents the normalized distance between the rows of SAMs from 2 different systems referred to as sys (E.g. ASR and SER) and are computed using Equation 5.1, where l and h refers to the transformer layers (= 12) and attention heads (= 12) in Wav2vec2.0 respectively.

$$Dis_{l,h} = \frac{\sum_{i=0}^{N} BC(A^{sys}_{l,h,i}, A^{sys}_{l,h,i})}{N} \tag{5.1}$$

In Figure 5.5, subplots (a) and (b) reveal that the attention behaves similarly in the initial layers, as indicated by the low distance values depicted in the plots. However, it is in the last few layers where the attention mechanism becomes more task-specific, as evidenced by the higher distance values. In Figure 5.5, subplot (a), when comparing ASR100 and SER, we observe that in the last layer, some attention blocks exhibit high distance values, while others have lower distance values. In contrast, in Figure 5.5, subplot (b), when comparing

ASR960 and SER, we see that the last layer attention heads have consistently high distance values. This observation might explain the results in Table 5.1, where ASR100 yields better results compared to ASR960 and even the PT network. It suggests that some attention heads in ASR100 focus on phonetic information, while others concentrate on emotion-related information. For ASR960, the attention does not correspond well to emotional information, as indicated by the high distance values. Upon further fine-tuning of these ASR(X) networks for the ER task (ASR(X)→SER), we observe that the attention heads in the last layer begin to behave similarly to those in the SER network. This is evident from the lower distance values in Figure 5.5, rightmost subplots for both (a) and (b). It is worth noting that even after adapting the network for the ER task some intermediary layers (e.g., 7, 8) still exhibit high distance values, showing the information was not restored completely; this requires further probing.

5.5 Summary

Our study explored information recovery potential of SFMs using SER and ASR tasks as the target and intermediary tasks, respectively. Initially, the evaluation of intermediary network representations for the target task uncovered an inverse relationship. As the network excelled in the ASR task, it exhibited a decline in SER discrimination properties. However, fine-tuning the intermediary networks for the target task successfully recovered SER information, achieving performance levels comparable (similar) to a network directly tuned for the target task. Despite similar overall performance, we identified disparities in decision-making capabilities between the networks (SER and ASR(X)→SER). Future investigations will explore information recovery using diverse tasks, and the interplay between different learning rates. Additionally, we aim to conduct a layer-wise analysis with different SFMs to further enhance our understanding of information recovery in these systems.

6 Model adaptation for Parkinson's Disease detection from speech

Having established in the preceding chapters that Speech Foundation Models (SFMs) learn powerful and transferable representations for paralinguistic tasks such as emotion recognition, we now shift focus to a more challenging frontier, modeling pathological speech. While our earlier investigations showed that fine-tuning SFMs can enhance emotion recognition performance, those findings were based exclusively on typical, healthy speech. In contrast, this chapter examines how these same models behave when confronted with atypical speech, marked by reduced prosody, degraded articulation, and voice quality changes as observed in individuals with Parkinson's Disease (PD). This shift: from modeling affect in healthy speakers to detecting pathology-related deviations, examines how well SFMs generalize and adapt to clinically relevant, low-resource scenarios.

PD, a neurodegenerative disorder caused by the progressive loss of dopaminergic neurons *(Hornykiewicz, 1998)*, often results in speech impairments such as reduced voice quality, monotonicity, and difficulty in articulation *(Baker et al., 1998; Rusz et al., 2013)*. Speech analysis offers a non-invasive, cost-effective approach for automatic PD detection, motivating the development of systems to reduce the time and effort required for clinical assessments of PD-related speech disorders, known as dysarthria *(Ngo et al., 2022; Virmani et al., 2022)*.

Traditional approaches to dysarthria detection have relied on handcrafted acoustic features such as jitter, shimmer, formants, glottal parameters, and Mel-Frequency Cepstral Coefficients (MFCCs) *(Bocklet et al., 2013; Prabhakera and Alku, 2018; Hawi et al., 2022)*. With the rapid progress of Deep Learning (DL) *(LeCun et al., 2015)*, research has increasingly focused on applying DL techniques to automatic pathological speech detection *(Gupta et al., 2016; Cummins et al., 2018)*. Recent studies have explored different architectures and speech representations for dysarthria detection: for example, Convolutional Neural Networks (CNNs) *(Vásquez-Correa et al., 2017; Janbakhshi et al., 2021; Kodrasi, 2021)*, Long Short-Term Memory (LSTM) networks *(Bhati et al., 2019; Bhattacharjee et al., 2021; Joshy and Rajan, 2021)*, and convolutional-recurrent hybrids combining 1D-CNNs with LSTMs for classifying Parkinson's disease (PD) versus healthy controls (HC) *(Rios-Urrego et al., 2022)*. The following authors: Garcia et al. (2017) and Moro-Velazquez et al. (2020) utilized latent features such

as i-vectors *(Dehak et al., 2010)* and x-vectors *(Snyder et al., 2018)*, originally developed for speaker verification and identification, to differentiate PD from HC. Wodzinski et al. (2019) and Karaman et al. (2021) highlights the limited availability of pathological speech data for training deep neural networks and thus employed a transfer learning approach for PD detection. Both studies utilized the ResNet architecture *(K. He et al., 2016)*, pretrained on the ImageNet corpus *(Russakovsky et al., 2015)*. Wodzinski et al. used the pretrained features for classification, while Karaman et al. fine-tuned the network for PD detection. Although the pretrained network was meant for image classification both the studies demonstrated the effectiveness of transfer learning for the PD detection task.

The scarcity of diverse, publicly available pathological speech data remains a challenge. However, the success of SFMs *(A. Mohamed et al., 2022)* in various speech-related downstream tasks has led to increased interest in leveraging transfer learning for pathological speech analysis. Recent studies *(Wagner et al., 2023; Amiri and Kodrasi, 2024; Escobar-Grisales et al., 2024; Wiepert et al., 2024)* highlight the effectiveness of SFMs, particularly wav2vec2.0-base *(Baevski et al., 2020)*, in encoding different speech pathologies.

When using SFMs, there are two primary approaches: freezing them to serve as feature extractors or fine-tuning/adapting them for downstream tasks, as illustrated in Figure 6.1 (a) and Figure 6.1 (b), respectively. Studies have demonstrated that when using SFMs as feature extractors, each layer captures distinct speech-related information *(Pasad et al., 2023; Chowdhury et al., 2024)*, making layer selection an advantageous approach for the task. At the same time, full fine-tuning/adaptation of SFMs for pathological speech is still underexplored. In this chapter, we focus on the PD detection task utilising SFMs and we: (a) propose a methodology for layer selection in SFMs, (b) investigate the effectiveness of full-finetuning on SFMs, and (c) investigate LoRA-based adaptation approach for PD detection, utilizing parameter-efficient fine-tuning (PEFT) via the Low-Rank Adaptation (LoRA) method *(Hu et al., 2022)*. While LoRA has been explored in various speech processing tasks *(T. Feng and Narayanan, 2023; W. Liu et al., 2024; Song et al., 2024)*, its application to pathological speech detection remains unexamined.

This chapter investigates how best to adapt large scale SFMs to the data-scarce and acoustically heterogeneous domain of Parkinsonian speech. By comparing layer selection strategies, full fine-tuning, and LoRA based parameter-efficient methods, we aim to uncover the trade-offs between efficiency, generalization, and representational robustness. Beyond the immediate goal of accurate PD detection, this exploration also probes a deeper question: Can the same architectures that excel at modeling expressive, healthy speech be repurposed to recognize and characterize the subtle degradations of neurodegenerative speech?

The rest of the chapter is organised as following, Section 6.1 introduces the methods investigated in this study. Section 6.2 outlines the dataset used and the experimental setup, Section 6.3 presents the experimental results and the analysis, finally Section 6.4 concludes the chapter.

6.1 Methods investigated

6.1.1 Cross-validation based layer selection

We propose a simple approach, (i) extract the representations of speech utterances from different layers of the SFMs (ii) train an auxiliary classifier for the task-on hand and evaluate it only using the cross-validation set (iii) use the classifier's accuracy on the cross-validation set as an indicator of the layer's effectiveness in learning task-related properties. (iv) select the best performing layer on the cross-validation set for evaluation on the test set.

6.1.2 Fine-tuning/Adaptation

Fine-tuning/adaptation involves updating all model parameters, including those of the foundation model (upstream model) and the classifier block (downstream model), to tailor the network for the target task, as shown in Fig 6.1(b).

6.1.3 Low-Rank Adaptation

LoRA is a PEFT technique, proposed to efficiently adapt large FMs to specific domains or downstream tasks. Consider $W_0 \in \mathbb{R}^{d\times k}$ to be the pre-trained weight matrix, LoRA replaces model weights with low-rank matrix decomposition as follows, $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and rank $r \ll min(d,k)$. During training W_0 is kept frozen while A and B are trainable parameters as shown in Figure 6.1(c) . This strategy significantly reduces the number of trainable parameters *(Hu et al., 2022)*.

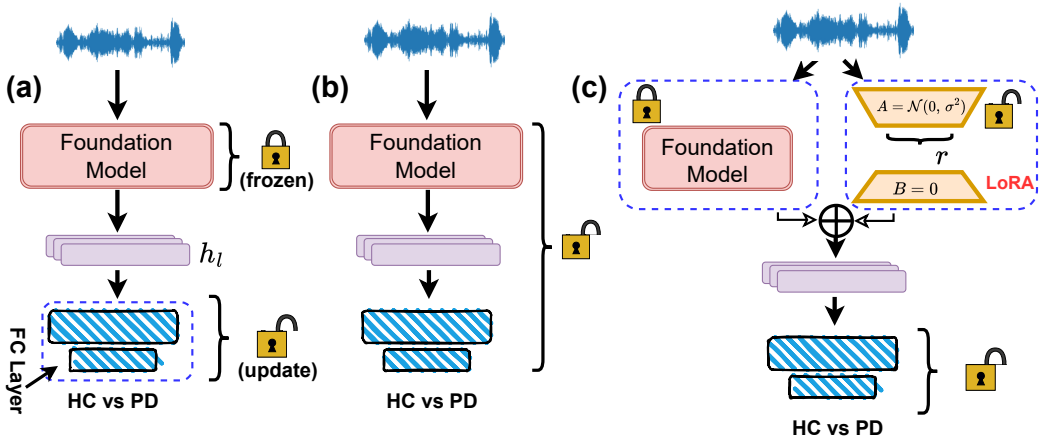


Figure 6.1: Figure depicting different training methodologies: (a) Linear probing, (b) Fine-Tuning, and (c) Fine-Tuning using LoRA ; 'h_l' and 'FC Layer' refer to frame-level embeddings from layer 'l' and the fully connected layer, respectively; HC= healthy control and PD= Parkinson's disease.

6.2 Experimental Setup

6.2.1 Dataset and Protocol

We consider PC-GITA corpus *(Orozco-Arroyave et al., 2014)* for the PD detection task. The corpus consists of 100 participants, divided equally between 50 Parkinson's disease (PD) patients and 50 healthy controls (HC), all native Spanish speakers from Colombia. The two groups are balanced in terms of age, gender, and education level. Each participant contributed 10 sentences and a phonetically balanced text, providing an average of 55.5 seconds of speech data per participant. Originally the speech data was recorded at a sampling frequency of 44.1 kHz, we downsampled the recordings to 16 kHz for our study.

For our investigation, we resort to a stratified 10-fold speaker-independent cross-validation evaluation. At each fold, 80%, 10%, and 10% of the data is used for training, cross-validation, and testing, respectively. Following the previous work, we report the performance of our classification systems in terms of accuracy, F1-score, sensitivity (correct classification rate for PD), and specificity (correct classification rate for HC). The final performance is the mean and standard deviation of classification metric values obtained across 10 folds of the test set. It is worth noting that with this protocol, we acquire approximately 1 hour and 15 minutes of speech data per fold for training.

6.2.2 System description and Configurations

(a) Handcrafted features: For Baseline we utilize knowledge-based feature representations provided with openSMILE toolkit *(Eyben et al., 2010)*. We utilize ComParE$_{LLD\times F}$, which consists of 6,373 static turn-level features derived from computing functionals (statistics) over (65 + 65) LLD contours. Additionally, we conducted experiments with eGeMAPS$_{LLD\times F}$, which includes 88 static turn-level features obtained from functionals computed over 23 LLD contours.

We use support vector machine (SVM) as a classifier for the handcrafted feature-based pipeline. SVM performance was optimized by doing grid search for hyperparameters such as the kernel, kernel width (γ) and soft margin constant (C) using the cross-validation set.

(b) SFMs: In this, we evaluate three SFMs for the PD detection task. These models were carefully chosen to facilitate a detailed comparison of different training paradigms, including self-supervised and weakly supervised learning. We also investigate the effect of monolingual versus multilingual pretraining, which is particularly relevant given that the PC-GITA corpus comprises Spanish speech recordings.
Wav2vec2.0-base *(Baevski et al., 2020)*, hereafter referred to as W2V2, utilizes a self-supervised learning approach, combining contrastive learning with masking techniques. It comprises 12 transformer encoder layers, 768-dimensional hidden states, and 8 attention heads, amounting to 95 million parameters. Pretraining of network was conducted on 960 hours of English audio

using the LibriSpeech corpus.
XLSR *(Conneau et al., 2021)* is a multilingual variant of Wav2vec2.0 with 24 transformer encoder layers, 1024-dimensional hidden states, and 16 attention heads. The model is pretrained on 53 languages and consist of 315M parameters.
Whisper-small *(Radford et al., 2023)* henceforth denoted as whisper, pretrained in a wealky-supervised fashion. The model comprises 12 encoder and 12 decoder blocks, each with 12 attention heads and a 768-dimensional hidden state, totaling 244 million parameters. Whisper is multilingual, and was trained on approximately 680,000 hours of weakly-supervised speech data sourced from the internet.

All the SFMs were retrieved from HuggingFace. The frame level representation derived from the SFMs were mean pooled and then fed to the classifier head consisting of one hidden layer with 256 nodes and output layer of 2 nodes corresponding to the number of classes, 2 in this case (HC and PD). The output layer had softmax activation, while the hidden layer had ReLU activation. The networks were trained using cross-entropy loss with Adam optimizer. The batch size was set to 4, with gradient accumulation configured at 8, and the seed value was set to 1337. When probing the network for the downstream task (Figure 6.1(a)), the learning rate was set to 1×10^{-4}. For fine-tuning (Figure 6.1(b)), the learning rate was adjusted to 1×10^{-5}. It is worth emphasizing that during fine-tuning, the CNN-based encoder blocks were kept frozen for all SFMs, while only the transformer encoder blocks were fine-tuned. In the case of whisper, the 12 encoder layers were used to generate speech representations, while the decoder layers were excluded. Lastly, for fine-tuning with LoRA (Figure 6.1(c)), we performed experiments with rank (r) value across the set {4, 8, 16} for each SFMs, reporting the best results here.

6.3 Results and Analysis

6.3.1 System performance

Figure 6.2 presents the accuracy trends for the three SFMs based on the layer-wise analysis performed on the cross-validation set on 10 folds. The results reveal that the best-performing layer for W2V2 is Layer 10, which aligns with findings from previous studies *(Wagner et al., 2023; Amiri and Kodrasi, 2024)*. For XLSR, Layer 16 shows the highest performance, while Whisper's optimal layer is Layer 12. Notably, XLSR exhibits minimal fluctuation in performance across the mid layers, with the last layer performing the worst.

Table 6.1 G-2 presents the test set outcomes for the layers selected using the cross-validation set. W2V2 and XLSR achieves higher accuracy, while XLSR and Whisper show better sensitivity scores. Comparing layer selection to fine-tuning, shown in Table 6.1 G-3, we observe a slight accuracy improvement for XLSR and whisper, but a decrease for W2V2. Additionally, the sensitivity score for XLSR increases after fine-tuning, while it decreases for W2V2 and remains unchanged for Whisper. When comparing layer selection results with LoRA adaptation in

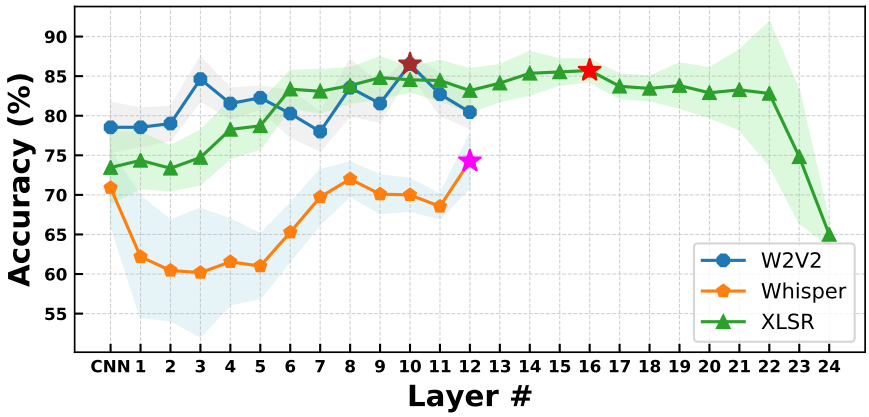


Figure 6.2: -•- on curves depicts mean of classification accuracy over 10 folds on validation-set at every layer. Best Accuracy (on validation-set data): -⋆- W2V2: 86.53%, -⋆- Whisper: 74.27%, and -⋆-XLSR: 85.72% from layers 10, 12 and 16 respectively.

Table 6.1 G-4, we observe a 4% absolute gain in whisper's accuracy. Furthermore, the results indicate that SFM-based classification whether using layer selection or fine-tuning, consistently outperforms the handcrafted features, reported in Table 6.1 G-1. Notably, we obtain the best-performing accuracy of 85.00% with whisper finetuned using LoRA with rank value 16 (in Table 6.1 G4). This surpasses what has been previously reported in the literature using a similar train test protocol for PC-GITA- 83% *(Janbakhshi et al., 2021)* and 82.6% *(Amiri and*

Table 6.1: *Comparison of different feature representations for PD vs. HC classification results on test-set, averaged over 10-folds on PC-GITA. (·) indicates the standard deviation. "Param." indicates network's trainable parameters for respective systems.*

Feature representation	Param.	Dim.	Accuracy ↑	F1-score ↑	Sensitivity ↑	Specificity ↑
G-1: Handcrafted Features						
ComParE$_{LLD\times F}$	–	6373	77.60(7.1)	77.62(7.1)	76.90(10.2)	79.16(12.3)
eGeMAPS$_{LLD\times F}$	–	88	76.44(6.3)	75.98(6.4)	75.45(11.8)	77.45(13.3)
G-2: Cross-validation layer selection for SFMs (selected layer)						
W2V2 (L 10)	197K	768	83.54(5.6)	83.73(5.3)	84.72(9.8)	82.36(13.4)
XLSR (L 16)	263K	1024	83.72(8.3)	84.12(8.1)	86.12(10.6)	81.27(13.2)
WHISPER (L 12)	197K	768	81.09(8.6)	81.93(8.0)	86.00(10.7)	76.18(12.1)
G-2.2: Combining decisions from all layers of SFMs						
W2V2	197K	768	84.27(6.5)	84.54(6.2)	85.82(9.4)	82.73(12.5)
XLSR	263K	1024	83.91(7.1)	84.45(6.7)	86.73(8.6)	81.09(12.9)
WHISPER	197K	768	75.73(9.6)	74.27(9.9)	70.73(12.7)	80.73(15.5)
G-3: Fine-tuning SFMs						
W2V2-FT	90.4M	768	80.90(8.3)	80.09(9.7)	79.27(15.6)	82.54(13.1)
XLSR-FT	311M	1024	84.72(7.8)	85.51(7.2)	89.45(8.6)	80.00(12.8)
WHISPER-FT	87.2M	768	83.53(8.3)	83.91(8.3)	86.06(11.5)	81.01(14.4)
G-4: Fine-tuning SFMs with LoRA adapters (rank)						
W2V2 (R 4)	862K	768	83.09(7.2)	83.06(7.4)	83.81(12.2)	82.36(13.1)
XLSR (R 4)	2M	1024	83.09(7.5)	83.30(7.3)	84.90(12.4)	81.27(15.4)
WHISPER (R 16)	2.9M	768	85.00(9.6)	85.34(8.9)	86.36(10.1)	83.63(15.2)

Kodrasi, 2024). When compared to previous work *(La Quatra et al., 2024)* which investigated PD classification on PC-GITA using only fine-tuning, our study presents a comparative analysis, demonstrating that selecting the appropriate layer and using PEFT-based fine-tuning can achieve performance comparable to full fine-tuning.

6.3.2 Analysis

Layer selection analysis: To better analyse our cross-validation layer selection scheme, we computed the layer-wise accuracy for the test set similarly to what we did for cross-validation data. Figure 6.3 showcases the accuracy trend for test set. We see some common layer-wise accuracy trend for both test and cross-validation data. For W2V2, layer-3 achieves an accuracy of 85.36% on the test set. When using cross-validation for layer selection, layer 10 is chosen, resulting in a slightly lower test accuracy of 83.54%. This small difference indicates that performance remains relatively stable across layers. If observed for cross-validation set (in Figure 6.2) next best pick would have been layer-3. For whisper the trend remains the same with lower layer not performing well, and the last layer (layer-12) being the best pick for both validation and test set. This outcome is reasonable, as whisper is trained for speech recognition (SR) task and last layer being task specific for SR could pick up on the cues of atypical PD speech. For XLSR, cross-validation selects layer-16, which achieves a test accuracy of 83.72%. However, layer 15 attains a higher test accuracy of 87.06%. This might be an anomaly, as observed from the peak at layer-15 in Figure 6.3. Otherwise, for XLSR the accuracy trend remains the same for the test and the cross-validation data, that is the middle layer yields better results, and for the last layer a drastic drop in accuracy is observed. This analysis showcases that cross-validation layer selection scheme generalises well to the test set. We also analyze the impact of combining decisions from all layers using a majority voting strategy. Table 5.1 (G2.2) presents the results, where no significant performance gain is observed, with

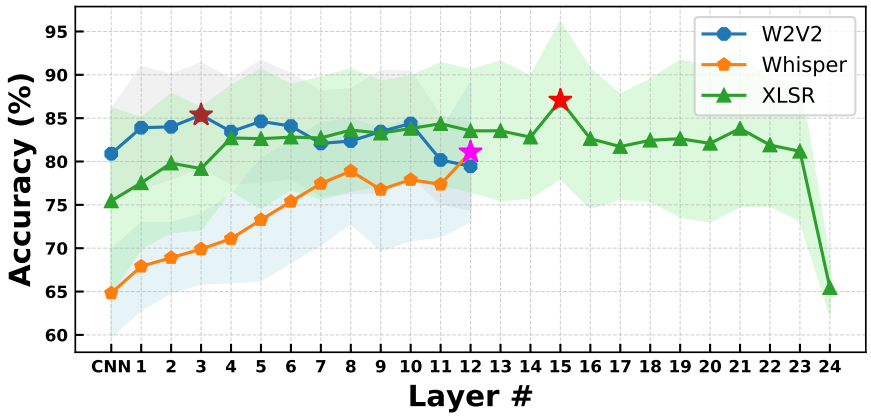
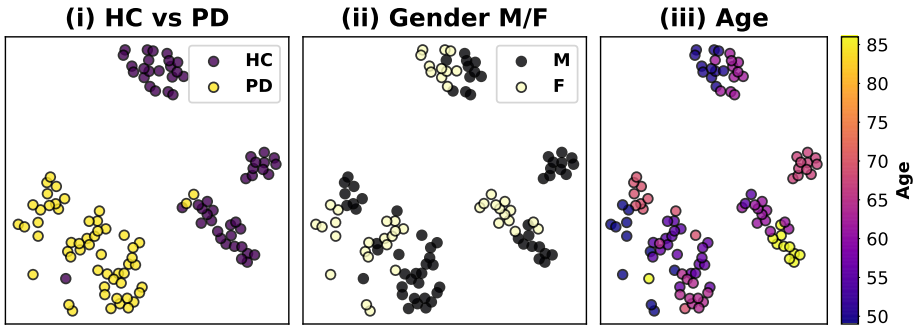


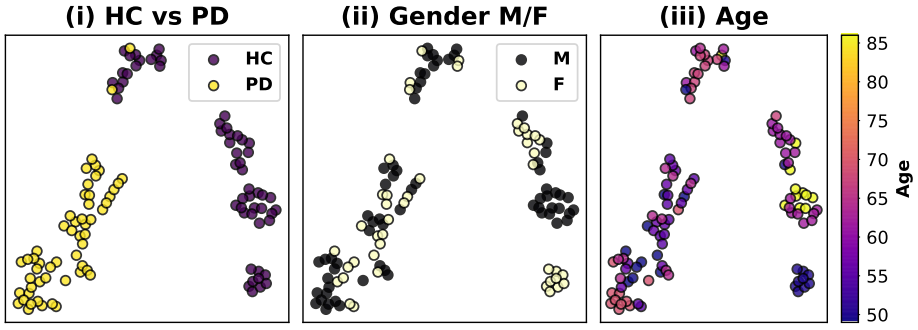
Figure 6.3: -•- on curves depicts mean of classification accuracy over 10 folds on test set at every layer. Best Accuracy (on test set data): -⋆- W2V2: 85.36%, -⋆- Whisper: 81.09%, and -⋆-XLSR: 87.06% from layer 3, 12 and 15 respectively

a noticeable drop in Whisper's performance. These findings, along with the trends shown in Figure 6.2 and 6.3, suggest that selectively choosing layers for combining decisions may be more advantageous.

Embedding space analysis: We generate embeddings from the test set data of one fold, consisting of 110 utterances from 10 speakers (5 HC and 5 PD), with 3 males and 2 females in each group. For embedding generation, we select two systems: XLSR (L 16) with 83.72% accuracy, and Whisper (R 16) fine-tuned using LoRA, achieving 85% accuracy. To visualize the embedding space, we use t-SNE plots, as shown in Figure 6.4. For XLSR (L 16) and Whisper (R 16), we observe distinct clusters for PD and HC. When these clusters are color-coded by gender and age, we notice distinct and systematic clustering, particularly within the HC set. This indicates that these networks might also be capturing speaker identity information in the process of PD detection.



(a) XLSR (L16)



(b) WHISPER (R 16)

Figure 6.4: *t-SNE plot of the last layer embedding space from selected systems, using 110 utterances from 10 speakers (5 HC and 5 PD) in the test set. Data points are color-coded to represent: (i) HC vs PD, (ii) Gender (M and F), and (iii) Age.*

6.4 Summary

In this work, we investigate the adaptability of SFMs for PD detection, focusing on how best to leverage their representational power under data-scarce, clinically atypical speech conditions. We propose a cross-validation-based layer selection methodology and compare its effectiveness against full fine-tuning of the models. Furthermore, we introduce, for the first time, a LoRA-based fine-tuning strategy for PD detection. Our findings reveal that the layer selection approach attains performance comparable to full fine-tuning while being considerably more parameter efficient. Interestingly, LoRA adaptation applied to Whisper surpasses the layer selection method likely because Whisper's pretraining for speech recognition allows LoRA-based fine-tuning to better capture the articulatory and prosodic deviations characteristic of PD speech. These insights open promising directions for future research, particularly in adapting ASR-pretrained SFMs for clinical speech analysis and early detection of neurodegenerative conditions.

7 Speech-based analysis of depression comorbidity in Parkinson's Disease

Paralinguistic analysis often explores human speech through two temporal lenses: transient 'states', which capture short-lived affects like emotions, and stable 'traits', which reflect long-term characteristics or pathologies such as Parkinson's Disease (PD). The preceding chapters examined these aspects separately, first by modeling emotional states, and later the pathological trait of PD from speech segments. This chapter bridges these two perspectives by investigating a clinically significant intersection: modeling an affective state (depression) within a neurological trait (PD). This study, therefore, has a dual objective: first, to understand how affective and motor symptoms acoustically intertwine, and second, to investigate the capacity of current paralinguistic modeling neural representations to disentangle these complex state-trait interactions in a comorbid condition.

Depressive disorder is a common mental health condition that affects an estimated 5% of the global adult population, according to the World Health Organization *(World Health Organization, 2023)*. It is associated with a range of emotional, cognitive, and behavioral symptoms *(Scibelli et al., 2018)*, such as persistent low mood, anxiety, and slowed motor activity, and in some cases may lead to suicidal thoughts *(World Health Organization, 2025)*. Diagnosis typically depends on clinical interviews, which can vary in consistency and may contribute to heterogeneous assessments *(Burdick et al., 1983; Landau, 2008; Cummins et al., 2015)*. This highlights the relevance of developing reliable automated approaches to support the screening of depression across its diverse manifestations.

In the speech community, numerous studies have explored the use of acoustic features for detecting depression, demonstrating the potential of vocal indicators *(Nilsonne et al., 1988; France et al., 2000; Moore II et al., 2007; Cummins et al., 2011; Stasak et al., 2016; Zahid et al., 2020)*. These studies typically involve statistical analysis of features such as fundamental frequency (F0), intensity, and spectral properties, suggesting their relevance as possible markers for depression screening. In recent years, artificial neural network (ANN)-based methods have become increasingly popular for this task, showing promising results in capturing speech patterns linked to depressive states *(X. Ma et al., 2016; L. He and Cao, 2018; Dubagunta et al., 2019; W. Wu et al., 2023)*. While these methods often outperform traditional handcrafted

approaches in terms of accuracy, they generally lack interpretability and are constrained by the availability of sufficient labeled data for training *(Favaro et al., 2023)*.

Although the field has seen substantial advances, accurately detecting depression from speech remains difficult due to the diverse manifestations of depression. An area that remains relatively underexplored is the study of depression when it co-occurs with neurological conditions such as PD or Alzheimer's disease *(H. Lee and Lyketsos, 2003)*. Aarsland et al. (2012) emphasize that while PD is primarily known for its motor impairments, it also encompasses a wide range of non-motor symptoms. Among these, depression is one of the most common, affecting approximately one-third of individuals with PD. Notably, depression in PD is often persistent and may even manifest during the prodromal phase *(Dušek et al., 2019)*.

The coexistence of depression with neurological disorders such as PD adds complexity to the depression diagnosis, and often results in inadequate treatment *(Hussain et al., 2020)*. Despite growing interest in speech-based mental health assessment, the detection of depression from atypical speech has received limited attention. To date, only a few studies have specifically addressed this challenge. Ozkanca et al. (2019) were among the first to investigate depression screening in PD patients using brief 10-second phoneme recordings. Their approach relied on handcrafted acoustic features combined with standard machine learning classifiers, demonstrating a promising relationship between vocal patterns and depression in PD. However, the study was limited in scope to isolated phoneme-level speech. In contrast, P. Pérez-Toro et al. (2021), P. A. Pérez-Toro et al. (2022), and P. A. Pérez-Toro et al. (2023) focused on longer speech samples, analyzing monologues from individuals with PD and Alzheimer's disease. They applied transfer learning strategies by first modeling affective states using ForestNet *(Rodríguez-Salas et al., 2020)* in the valence-arousal space and then fine-tuning the models for depression detection. While their work demonstrates the potential of using emotion modeling as an intermediate task, it still leaves open questions regarding direct depression classification from atypical, disorder-specific speech. Together, these studies highlight a notable gap in the literature: the limited investigation into depression detection from neurologically atypical speech signals, where both motor and affective symptoms may alter vocal expression.

In this study, we first examine the efficacy of using interpretable handcrafted acoustic features, commonly successful in general depression detection, to specifically detect depression in individuals with PD. Using two distinct corpora, we analyze and compare the acoustic characteristics of individuals with comorbid PD-Depression against those with depression alone, aiming to isolate the factors that confound accurate depression classification. Furthermore, we introduce a feature filtering method based on the Point Biserial Correlation Coefficient (PBCC) to identify the most discriminative features for this complex task. To provide clarity in our analysis, and following the acoustic descriptor definitions in Eyben (2015), we categorize the acoustic features into three main groups based on their underlying physiological and physical properties: vocal source-related features (pitch, loudness, and voice quality), vocal tract-related features (formants and spectral properties), and global-related features (overall

energy and rhythm). Finally, for completeness and to further explore the challenges of modeling this comorbidity, we expand our investigation by incorporating representations derived from state-of-the-art Speech Foundation Models (SFMs), using the cross-validation-based layer selection approach previously developed in Chapter 6 for the PD detection task.

The remainder of the chapter is structured as follows: Section 7.1 describes the datasets used in the study. Section 7.2 outlines the methods explored, while Section 7.3 details the feature representations and experimental setup. Section 7.4 presents the experimental results, followed by a discussion and analysis in Section 7.5. Finally, Section 7.6 concludes the chapter.

7.1 Dataset

Depression in Parkinson's disease (PD-D) *(P. Pérez-Toro et al., 2021)*: The dataset comprises speech recordings from 60 Colombian Spanish speakers, including 25 Depressive Parkinson's Disease (D-PD) patients and 35 Non-Depressive Parkinson's Disease (ND-PD) patients. Each participant was asked to deliver a monologue describing their daily routine. Following the recordings, a neurologist assessed their neurological condition using the Movement Disorders Society - Unified Parkinson's Disease Rating Scale (MDS-UPDRS) *(Goetz et al., 2008)*, which is the standard tool for evaluating the neurological state of PD patients. The initial section of the MDS-UPDRS includes an item that evaluates depression based on daily routine activities, assigning scores from 0 to 4. Participants scoring above zero were categorized as D-PD, while those with a score of zero were identified as ND-PD. The average length of the monologues is 84±34 seconds for D-PD and 80±37 seconds for ND-PD, resulting in a total dataset duration of approximately 4892 seconds. Speaker 52 was excluded from the analysis due to recording issues.

Distress analysis interview corpus - Wizard of Oz (DAIC-WOZ) *(Gratch et al., 2014)*: The dataset includes audio-visual interviews from 189 participants who were evaluated for psychological distress, amounting to 17 hours of audio data. Each participant was assigned a self-assessed depression score based on the Patient Health Questionnaire (PHQ-8) method *(Kroenke et al., 2009)*. For our experiments, we extracted only the participants' speech segments, using the time labels provided in the dataset to isolate their corresponding portions of the recordings. Sessions 318, 321, 341, and 362 were excluded from the training set due to time-labelling errors.

Table 7.1 summarizes the two datasets.

Table 7.1: Distribution of utterances used in the study, corresponding to each label.

Database	Content	Depressed patients	Not-Depressed patients	Total
DAIC-WOZ	English	42	100	142
PD-D	Spanish	24	35	59

7.2 Methodology

This study employs two distinct approaches, which are outlined below.

7.2.1 Feature selection for handcrafted acoustic descriptors

As a baseline approach, handcrafted features were initially extracted from the input audio signal and used as input representations for the classifier module to produce confidence scores (see solid arrows in Figure 7.1). This method provides a useful point of comparison with more advanced architectures.

Building upon the handcrafted-feature based approach, we introduce a feature selection step to refine the input representation. After extracting features from the raw audio signal, we apply the Point Biserial Correlation Coefficient (PBCC) to identify the most informative subset of features (see dashed arrows in Fig. 7.1). PBCC measures the strength of the linear association between a binary target variable (D/ND) and continuous-valued features, enabling the selection of features that are the most relevant for classification. The PBCC is defined as:

$$r_{pb} = \frac{M_D - M_{ND}}{s_n} \sqrt{\frac{n_D n_{ND}}{n^2}}$$

where M_D and M_{ND} denote the mean feature values for the Depressive (D) and Non-Depressive (ND) classes, respectively; n_D and n_{ND} are the number of samples in each class; n is the total number of samples; and s_n is the standard deviation of the feature values across all samples.

To begin with, a range of correlation thresholds is defined. For each threshold value, the PBCC is calculated between every feature and the target labels in the training set. Features that exceed the given threshold are retained, as they demonstrate a stronger linear association with the class labels and are considered more discriminative.

Next, a Gradient Boosting (GB) classifier is employed to evaluate the predictive performance of the selected feature subsets. Thresholds ranging from 0.06 to 0.6, with increments of 0.2, are tested to determine which subset yields the highest classification accuracy.

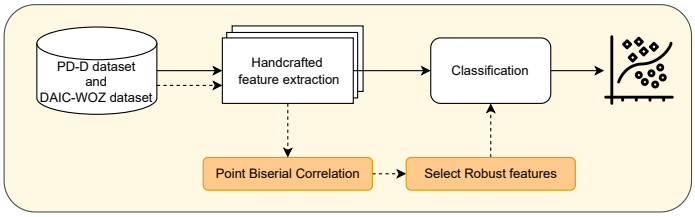


Figure 7.1: Proposed methodologies representation: conventional approach (solid arrows) and PBCC-based approach (dashed arrows).

7.2.2 Cross validation based layer selection for SFMs

The proposed architecture consists of two main components: an upstream SFM encoder and a downstream classification module. The upstream encoder comprises a stack of N transformer layers, each incorporating self-attention mechanisms. The downstream module includes an average pooling layer followed by a fully connected multilayer perceptron (MLP).

To determine which encoder layer produces the most task-relevant representations, we adopt the following procedure as established in Chapter 6: (i) Representations are extracted from each layer of the SFM encoder using the input speech utterances. (ii) The downstream network is trained using these representations to perform the required task, and its performance is evaluated exclusively on the cross-validation set. (iii) The accuracy on the cross-validation set serves as a proxy for the effectiveness of the representations in capturing task-specific information. (iv) The layer that achieves the highest cross-validation accuracy is selected for final evaluation on the test set.

The training procedure for the proposed model is as following: Let X represent the input speech signal and $c \in \{D, ND\}$ the corresponding class label. The objective is to estimate the posterior probability $P(c|X)$

$$\begin{aligned} H_L &= \mathrm{ENC}_L(X;\theta_{e_L}), \\ P(c|X) &= \mathrm{softmax}(\mathrm{MLP}(\mathrm{Pool}(H_L);\theta_m)) \ \forall c \end{aligned}$$

Here, $H_L \in \mathbb{R}^{T\times D}$ denotes the sequence of embeddings from the L-th layer of the encoder, where T is the sequence length and D the embedding dimension. $\mathrm{ENC}_L(\cdot)$ is the encoder function for layer $L \in \{1,\ldots,N\}$ with parameters θ_{e_L}. The pooling function $\mathrm{Pool}(\cdot)$ aggregates the sequence of embeddings into a fixed-length representation, which is then passed through the MLP, parameterized by θ_m, followed by a softmax to produce the class probabilities.

The predicted label $\hat{c}$ is assigned by selecting the class with the highest posterior probability:

$$\hat{c} = \arg\max_c P(c|X)$$

Model training is guided by minimizing the cross-entropy loss $\mathscr{L}$ between the predicted distribution $P(c|X)$ and the ground truth label c:

$$\mathscr{L} = -\log P(c|X)$$

The encoder parameters θ_{e_L} are initialized using weights from a pretrained model and remain frozen during training. In contrast, the downstream parameters θ_m are randomly initialized and optimized during training.

7.3 Feature description and training protocol

In this section, we provide a brief overview of the handcrafted features and SFMs, along with the training protocol applied for each setup.

7.3.1 Handcrafted features

We use three widely recognized sets of knowledge-based handcrafted features (See Section 2.5.1): eGeMAPS *(Eyben et al., 2016)*, ComParE *(Schuller et al., 2013b)*, and DisVoice *(Orozco-Arroyave et al., 2018)*. The eGeMAPS set comprises 25 Low-Level Descriptors (LLDs) that capture essential acoustic properties, including frequency, energy, amplitude, and spectral characteristics. These descriptors are further processed using a set of statistical functionals, yielding a total of 88 features. ComParE provides a significantly larger feature set, comprising 6373 descriptors that include delta coefficients at the frame level and statistical functionals at the utterance level, all extracted via the openSMILE toolkit *(Eyben et al., 2010)*. DisVoice features, on the other hand, combine static representations related to phonation, articulation, and prosody into a unified feature vector. Further details and extraction procedures are available in Orozco-Arroyave et al. (2018).

For the classification task, we employed three machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB). Hyperparameters for each model were optimized using grid search in conjunction with 6-fold cross-validation, ensuring reliable and robust model selection.

To maintain consistency with prior work *(P. Pérez-Toro et al., 2021)*, we employed the Leave-One-Speaker-Out (LOSO) cross-validation protocol for the PD-D corpus. Specifically, for evaluating the k-th speaker, the model was trained on data from the remaining $k-1$ speakers. For the DIAC dataset, evaluations were conducted on the development set, as the official test set was reserved for the AVEC 2016 challenge Valstar et al., 2016.

7.3.2 SFM derived neural representations

In this work, we utilize two SFMs for the task of PD-D detection.

Wav2vec2.0-base *(Baevski et al., 2020)*, hereafter referred to as W2V2, adopts a self-supervised learning framework that integrates contrastive learning with masked prediction. The model architecture consists of 12 transformer encoder layers with 768-dimensional hidden states and 8 attention heads, totaling approximately 95 million parameters. It was pretrained on 960 hours of English speech from the LibriSpeech corpus.

XLSR *(Conneau et al., 2021)* is a multilingual extension of Wav2vec2.0, featuring 24 transformer encoder layers, 1024-dimensional hidden states, and 16 attention heads, resulting in a model

size of 315 million parameters. It was pretrained on speech data spanning 53 languages.

These models were deliberately chosen to facilitate a comparative analysis of the impact of monolingual versus multilingual pertaining, a relevant factor given that the PD-D corpus comprises Spanish-language recordings. SFMs were obtained from HuggingFace. The downstream head composed of a single hidden layer with 256 nodes and an output layer of 2 nodes, corresponding to the two target classes (Healthy Control or HC and Parkinson's Disease PD). The hidden layer employed ReLU activation, while the output layer used a softmax activation to produce class probabilities. Model training was performed using the Adam optimizer and cross-entropy loss. A batch size of 4 was used, with gradient accumulation set to 8 and the seed value was set to 1337. During probing for the downstream classification task, a learning rate of 5×10^{-4} was used for W2V2, and 1×10^{-4} for XLSR.

To evaluate the PD-D corpus using the SFM layer selection approach, we employed a 5-fold cross-validation strategy. Each fold consisted of speaker-independent training, validation, and test sets, comprising 35, 12, and 12 samples respectively.

7.4 System performance

This section reports the results and examines the classification performance associated with each feature category.

7.4.1 Handcrafted features

Table 7.2 summarizes the results obtained using the best-performing classifier- Gradient Boosting, (GB) for each feature set across both corpora considered in this study. In the DAIC-WOZ dataset, our proposed system substantially outperforms the baseline results reported in the AVEC 2016 challenge. These systems were based on the eGeMAPS feature set with an SVM classifier, achieving an improvement of approximately 51% in F1-score. While the conventional use of eGeMAPS yields an F1-score of 0.74, performance slightly drops to 0.69 following feature selection, likely due to the limited dimensionality of the feature set. Conversely, both ComPARE and $DisVoice$ show marked performance gains after applying feature selection.

For the PD-D corpus, our methods surpass the overall F1-score of 0.70 reported by Pérez-Toro et al. P. Pérez-Toro et al., 2021, who employed Valence and Arousal representations for classification. Among the conventional approaches, $DisVoice$ features initially exhibit the highest performance with an F1-score of 0.74. However, applying Point-Biserial Correlation Coefficient (PBCC)-based feature selection leads to significant improvements across all feature sets. Notably, ComPARE improves from 0.43 to 0.78 while using a reduced subset of just 186 dimensional feature.

Overall, these findings demonstrate that PBCC-based feature selection effectively enhances

Table 7.2: Classifiers' performance over the two datasets. *Dims* denotes the feature dimension; *Thr.* signifies the threshold set for feature selection; *D* and *ND* denote depressed and not-depressed patients, respectively; *O* is the unweighted average of *D* and *ND*.

			F1-score			Precision		Recall	
Features	**Dims**	**Thr.**	**O**	**D**	**ND**	**D**	**ND**	**D**	**ND**
DAIC-Woz									
Valstar et al. (2016)	88		0.49	0.41	0.58	0.26	0.94	0.88	0.42
Conventional approach									
eGeMAPS	88		**0.74**	0.62	0.87	0.90	0.78	0.47	0.97
ComPARE	6373		0.47	0.24	0.70	0.45	0.59	0.16	0.86
DisVoice	620		0.55	0.33	0.77	0.50	0.69	0.25	0.87
Feature-selection approach									
eGeMAPS	39	0.18	0.69	0.54	0.84	0.67	0.72	0.33	0.91
ComPARE	2756	0.18	**0.72**	0.65	0.80	0.62	0.78	0.33	0.87
DisVoice	184	0.20	0.65	0.47	0.83	0.80	0.73	0.33	0.96
PD-D									
P. Pérez-Toro et al. (2021)			0.68	-	-	-	-	-	-
Conventional approach									
eGeMAPS	88		0.54	0.40	0.69	0.53	0.61	0.32	0.79
ComPARE	6373		0.43	0.30	0.56	0.33	0.53	0.28	0.59
DisVoice	620		**0.74**	0.69	0.78	0.71	0.77	0.68	0.79
Feature-selection approach									
eGeMAPS	7	0.26	0.65	0.57	0.72	0.62	0.68	0.52	0.76
ComPARE	186	0.3	**0.78**	0.75	0.81	0.73	0.82	0.76	0.79
DisVoice	16	0.32	0.77	0.73	0.81	0.75	0.80	0.72	0.82

classifier performance, especially in identifying depressed (D) patients. The only exception is the slight decline observed for eGeMAPS in the DAIC-WOZ dataset. These results highlight the benefit of filtering out redundant or non-informative features, enabling the development of more compact, interpretable, and discriminative models.

7.4.2 SFM based features

Figure 7.2 illustrates the accuracy trends of the two SFMs based on layer-wise analysis conducted over 5-fold cross-validation. The results indicate that Layer 6 yields the highest performance for W2V2, while Layer 16 leads to optimal performance for XLSR. Table 7.3 presents

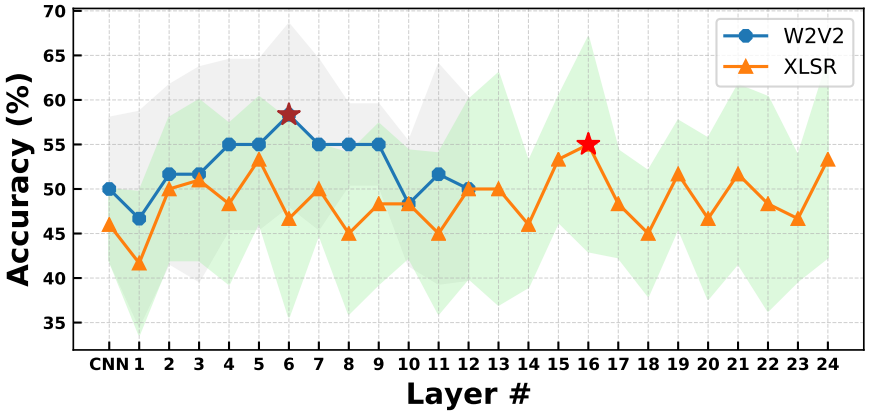


Figure 7.2: -•- on curves depicts mean of classification accuracy over 5 folds on validation-set at every layer. Best Accuracy (on validation-set data): -⋆- W2V2: 58.33%, and -⋆-XLSR: 55.00% from layer 6, and 16 respectively.

the test-set results on the PD-D corpus using the layers selected through the cross-validation-based layer selection strategy. The findings clearly indicate that the SFM-derived representations alone do not yield satisfactory performance for this task. Notably, the XLSR model pretrained on multiple languages including Spanish achieves an unweighted average F1-score of 40%, whereas the W2V2 model, pretrained solely on English, attains a higher score of 45%.

Table 7.3: SFM performance for the selected layer on test-set. *Dims* denotes the feature dimension; *D* and *ND* denote depressed and not-depressed patients, respectively; *O* is the unweighted average of *D* and *ND*.

		F1-score			**Precision**		**Recall**	
Features	**Dims.**	**O**	**D**	**ND**	**D**	**ND**	**D**	**ND**
				PD-D				
W2V2 (L 6)	768	0.45	0.19	0.70	0.43	0.58	0.12	0.88
XLSR (L 16)	1024	0.40	0.12	0.67	0.29	0.56	0.08	0.85

7.5 Result analysis and discussion

We now present a detailed analysis and interpretation of the results reported in the previous section.

7.5.1 Handcrafted features

Building on the classification results of knowledge based handcrafted features (shown in Section 7.4.1), we carried out a feature importance analysis by identifying the top 10 features according to their normalized importance scores, as determined by the best-performing classifier from the feature selection pipeline on both datasets. To facilitate interpretation, we organized these acoustic descriptors into three distinct categories based on the type of information they represent. Following the classification framework proposed by Eyben (2015), we divided Low-Level Descriptors (LLDs) into those related to the vocal source and those linked to the vocal tract. Additionally, we introduced a third category encompassing features that capture global properties of the speech signal, integrating information from both source and tract. Table 7.4 lists the selected features in descending order of importance, as assigned by the GB model. For reference, the descriptor names are provided, with the 'Index' column indicating the position of the feature in the openSMILE feature list, and the 'Group' column showing the assigned category.

The feature rankings in Table 7.4 reveal a distinct pattern in how the classifier identifies depression in speech from individuals with and without PD. In the DAIC-WOZ corpus, which involves classifying depressive versus healthy control subjects, the classifier places the greatest emphasis on source-related features—accounting for 6 out of the top 10. This is followed by a smaller number of vocal tract features (3 out of 10) and just one global acoustic descriptor. The prominence of source-related features suggests that vocal characteristics tied to the stability and quality of vocal fold vibrations such as: jitter, harmonic-to-noise ratio, and pitch variability are particularly informative for depression detection in a neurologically healthy population. These results align with clinical findings, such as those by Hollien (1980), who reported pitch alterations in depressed individuals, and Darby et al. (1984), who noted diminished pitch range and vocal intensity among depressed patients. In contrast, the feature rankings for the PD-D dataset indicate a broader spread of informative features. Notably, the presence of

Table 7.4: Feature ranking of GB trained on ComPARE for both DAIC-WOZ (left) and PD-D (right), using PBCC feature selection approach. Index column indicates the i -th feature from the 0-indexed feature list from the ComPARE header extracted from openSMILE

DAIC-WOZ			PD-D		
Index	**LLD name**	**Group**	**Index**	**LLD name**	**Group**
3861	logHNR	Source	142	Length L1	Global
4038	Jitter DDP	Source	88	RMS Energy	Global
5126	Py Sharpness	Source	1	Length L1	Global
1493	Spectral Harmonicity	Source	64	RMS Energy	Global
6077	Spectral Variance	Global	6067	Spectral Entropy	Vocal tract
6174	MFCC	Vocal tract	1245	Spectral Flux	Global
2365	audSpec	Vocal tract	1476	Py Sharpness	Source
4132	F0	Source	2925	Spectral RollOff 90.0	Vocal tract
4131	F0	Source	2991	Spectral Centroid	Global
1373	Spectral Skeweness	Vocal tract	3106	Spectral Kurtosis	Global

several spectral features such as- spectral entropy, spectral centroid, spectral roll-off, spectral flux, and spectral kurtosis reflects the vocal instability that is often observed in individuals with PD *(Hauptman et al., 2019)*. The observed reductions in spectral entropy and spectral centroid among depressed patients further highlight the nuanced interplay between emotional states and vocal quality *(Hussenbocus et al., 2015)*. Additionally, the feature Length L1, which captures aspects of voice quality and consistency, is impacted, pointing to irregular speech patterns likely caused by motor impairments. Overall, this distribution underscores the added complexity in detecting depression in PD patients due to the interplay between emotional and motor-related vocal alterations.

7.5.2 SFM based features

While handcrafted acoustic features enable interpretability and facilitate detailed analysis, such transparency is not readily achievable with representations derived from SFMs. Interestingly, contrary to earlier findings where SFM-based features have outperformed handcrafted ones for tasks like depression detection *(W. Wu et al., 2023)* and Parkinson's disease classification *(Favaro et al., 2023; Purohit et al., 2025a)*, this performance advantage diminishes in scenarios where depression coexists with Parkinson's disease. One possible explanation lies in the limitations of SFMs, which are typically pretrained on large-scale general speech corpora and may thus lack sensitivity to the subtle interplay between affective cues and motor deficits characteristic of this comorbid condition. The performances obtained further support this observation, for instance, when we analyse the mean sensitivity (correct classification rate for PD-D) and specificity (correct classification rate for PD-ND) scores across five folds, we observe a high specificity of 82% and a notably low sensitivity of 24% for W2V2 (L 6). This imbalance hints at the model's bias toward the majority class and its difficulty in capturing information related to depression within the pathological speech condition. Overfitting to the training data may also contribute to this limitation. To address these issues, future work will explore data augmentation and balancing strategies aimed at improving model robustness and sensitivity to underrepresented classes.

7.6 Summary

This study investigates automatic depression detection in individuals with Parkinson's disease (PD) using the PD Depression (PD-D) corpus. We explore two distinct feature types for this work: (1) interpretable, knowledge-based handcrafted acoustic representations, and (2) non-interpretable features derived from Speech Foundation Models (SFMs). For the handcrafted feature analysis, we not only evaluate their performance on the PD-D corpus but also study and compare acoustic descriptor patterns with those observed in typical depression-affected speech using the DAIC-WOZ corpus. The evaluation of SFM-derived features is focused exclusively on the PD-D dataset to assess their effectiveness in detecting depression within atypical speech associated with PD.

The classification results indicate that handcrafted feature-based approaches, such as those using the COMPARE feature set, can achieve an F1-score of 43% for depression detection in individuals with Parkinson's disease. Notably, performance improves substantially with the F1-score rising to 78% when redundant features are removed using a Point Biserial Correlation-based feature selection strategy. These findings highlight the critical role of feature selection in enhancing model robustness and discriminative power. In contrast, the performance of SFM-derived features remains subpar, even when representations are extracted from the optimal encoder layer identified via cross-validation layer selection methodology. The results indicate that these models tend to exhibit a bias toward the majority class, indicating limited effectiveness in capturing depression-related cues within the atypical speech patterns of individuals with Parkinson's disease.
The analysis of handcrafted feature ranking (see Section 7.5.1) highlights distinct acoustic profiles of depression in non-PD versus PD speech. In the DAIC-WOZ dataset, the classifier places strong emphasis on source-related features, highlighting their importance in detecting depressive states. This observation aligns with previous studies showing that vocal fluctuations, such as pitch variability, serve as reliable indicators of depression. Whereas, speech from individuals with PD exhibits a wider prominence of spectral features, reflecting the intricate speech alterations associated with the motor symptoms characteristic of the disorder.

Our findings suggest that the presence of Parkinson's disease complicates the automatic classification of depression, likely due to overlapping and interacting acoustic symptoms. The study demonstrates that depression manifests differently in the speech of individuals with Parkinson's compared to those without, emphasizing the need for tailored approaches in speech-based mental health assessment. Our results demonstrate that handcrafted features tend to yield better PD-D detection than SFM-derived representations.

One limitation of this study is the reliance on a low-resource and imbalanced dataset, where depression labeling is based solely on a non-zero score for the "Depressed mood" item in the MDS-UPDRS scale. We believe, a more robust labeling criterion would be necessary to enhance the reliability of the ground truth. Additionally, the dataset does not include information on medication status or speech-related subscores, both of which could provide valuable context for experimental design and interpretation of results. Addressing these limitations would require a separate dedicated data collection in the future.

8 Conclusions and future directions

In this chapter, we first summarize the key findings and then outline potential directions for future research.

8.1 Conclusion

The main focus of this work was on systematic exploration of the trade-offs, capabilities, and constraints across various speech modeling paradigms, spanning traditional handcrafted acoustic features, modern end-to-end deep learning, and large-scale Speech Foundation Models (SFMs). Within this context, the thesis addresses four research questions (RQs), initially focusing on paralinguistic states (speech emotion recognition or SER) and subsequently investigating paralinguistic traits (Parkinson's Disease or PD speech detection).

Conventional SER has relied on suprasegmental modeling, achieved through the extraction of hand-crafted low-level acoustic descriptors. Recent advances, however, favor utterance-level modeling, where speech representation such as Mel-Frequency Cepstral Coefficients (MFCCs) or long duration raw audio waveforms are processed directly by deep neural networks (DNNs) for SER. Contrary to these approachs, Chapter 3 introduced a novel approach to address RQ1: the feasibility of effectively learning emotion-discriminative information from short speech segments ($\approx$ 250 ms) using Convolutional Neural Network (CNN). Across multiple corpora (IEMOCAP, MuSe-Stress, and ExVo) and various emotion recognition tasks, our short-segment modeling achieved performance that was often comparable to or even better than traditional utterance-level systems. This approach demonstrated the effectiveness of short segment-level emotion encoding. We further showed that when handcrafted features were restricted to the same 250ms segment length, they yielded performance inferior to the end-to-end system, highlighting the benefit of the network directly learning optimal emotion discriminatory features from the raw data. The relevance signal analyses of our CNN revealed that the networks specifically emphasize emotion-relevant cepstral information. This suggested that fine-grained acoustic events act as robust carriers of affective information. This work showcased, for the first time, the efficacy of using very short temporal windows as powerful

indicators of emotional state.

In Chapter 4, we proposed a novel phonetically aware neural modeling framework to address RQ2 by evaluating the usefulness of phonetically informed representations for speech emotion recognition. The framework leveraged neural representations from networks trained or fine-tuned for phoneme or grapheme classification, enabling the extraction of phonetic information without requiring explicit transcriptions. This approach consistently outperformed traditional handcrafted acoustic features on benchmark English corpora. However, only limited additional gains were observed when compared to pre-trained SFMs, in line with later findings showing that such models already encode substantial phonetic information. Crucially, we also expose a trade-off between optimizing for linguistic content to make SFMs phonetically aware and preserving paralinguistic information (for SER); that is, ASR-oriented fine-tuning can attenuate emotion-relevant information. Subsequently, we showed in Chapter 5 that such paralinguistic information is recoverable via a second adaptation stage, albeit with the network converging to a different representational state than models optimized directly for emotion, evidence that "loss" under task transfer reflects reallocation rather than irreversible erasure.

While Chapters 3, 4, and 5 demonstrated the utility of SFMs for modeling paralinguistic states (emotions), Chapter 6 investigated RQ3: whether SFMs pretrained on healthy speech can be effectively utilized for modeling neurological traits, specifically in the context of PD speech analysis. In particular, we addressed the utility of SFMs for low-resource PD speech detection. We propose and validate parameter-efficient adaptation strategies, including a cross-validation-based layer selection methodology and the first application of Low-Rank Adaptation (LoRA) for PD detection. We found that the layer selection approach achieved performance comparable to full fine-tuning with significant parameter efficiency. Except, the application of LoRA adaptation to the Whisper model outperformed the layer selection approach. This finding suggests that models pretrained for speech recognition may inherently capture the subtle articulatory and prosodic deviations between PD and healthy speech, positioning them as efficient and suitable candidates for adaptation in PD analysis.

Finally, Chapter 7 of the thesis confronted the significant, under-explored challenge of detecting comorbid depression in Parkinson's Disease (PD). This study served as a complex test scenario for RQ4: assessing the robustness of neural representations in encoding a state (depression) within the context of a governing neurological trait (PD). Our key finding was that the acoustic manifestations of this comorbid depression are distinct from typical depression, as they are fundamentally confounded by PD speech symptoms. In this complex, low data-resource setting, we found that the large-scale SFM approaches failed, exhibiting high bias and poor sensitivity. Instead, a more traditional, interpretable approach using handcrafted features combined with a robust feature selection method proved significantly more effective, thus answering RQ4.

Collectively, this thesis advances the field by providing new approaches for fine-grained

emotion recognition and efficient methods for pathological speech analysis. It also offers a nuanced and critical perspective, demonstrating that while deep learning models are powerful, their limitations in complex, comorbid, and low-resource scenarios necessitate a pragmatic, task-aware approach, where knowledge based handcrafted representation remain not only relevant but, in some cases, superior.

8.2 Limitations and future directions

The findings, methods, and identified limitations from this thesis open up several promising avenues for future research.

Capturing multi-scale granularities: We established 250 ms as a viable segment length for modelling emotion information. It would be interesting to investigate a multi-scale architecture that simultaneously processes speech at different granularities (e.g., 250 ms for phonetic-level cues, 1-2 seconds for prosodic-level cues, and the full utterance for global context) and learns to fuse this information for robust emotion recognition.

Broader phonetic-class representation: For our work on phonetically aware neural representation it would be valuable to explore models pre-trained on broader phonetic classes (e.g., vowels, fricatives, nasals) to see if these more general articulatory-event embeddings offer a better trade-off between phonetic detail and emotional expression, as hinted at by earlier research.

Exploring auxiliary tasks for information recovery: Our framework for information recovery, used ASR as the intermediary task that "loses" paralinguistic information. This could be expanded to a matrix of intermediary tasks. For example, how does fine-tuning a pre-trained (PT) model for speaker identification ('PT → SpeakerID → SER') or even accent identification ('PT → AccentID → SER') affect the representation of emotion? This would create a comprehensive map of which paralinguistic tasks are synergistic and which are antagonistic.

Symptom-specific data augmentation: One of the bottleneck when dealing with pathological speech tasks is the small, imbalanced dataset. The most urgent future work is to explore robust data augmentation. That goes beyond simple noise addition and toward symptom-specific augmentation. This could involve, for example, using generative models (like Generative Adversarial Networks or Variational Autoencoders) trained to independently control acoustic parameters related to, for example, PD motor symptoms (e.g., tremor, dysarthria) or depression symptoms (e.g., F0 range, intensity), allowing for the synthesis of new, realistic pathological training samples.

Multimodal paralinguistic modeling: While this thesis focuses on speech as a modality, human communication is inherently bimodal, and future research should move toward multimodal frameworks that integrate visual dynamics and linguistic context. Beyond observable behaviors, incorporating involuntary physiological signals, such as heart rate variability, elec-

trodermal activity, and respiration, offers a promising direction. Fusing internal biological markers with external cues can enable more robust, accurate, and resilient paralinguistic systems.

Bibliography

Aarsland, D., S. Påhlhagen, C. Ballard, U. Ehrt, and P. Svenningsson, (2012). "Depression in Parkinson disease—epidemiology, mechanisms and management". in: Nature Reviews Neurology (cit. on p. 70).

Alku, P., J. Horáček, M. Airas, and A.-M. Laukkanen (2005). "Assessment of glottal inverse filtering by using aeroelastic modelling of phonation and FE modelling of vocal tract". In: *Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*. ISCA, pp. 73–76 (cit. on p. 10).

Amiri, M. and I. Kodrasi (2024). "Adversarial Robustness Analysis in Automatic Pathological Speech Detection Approaches". In: *Proc. of Interspeech*. ISCA, pp. 1415–1419 (cit. on pp. 60, 63 sq.).

Amiriparian, S., L. Christ, A. König, E.-M. Meßner, A. Cowen, E. Cambria, and B. W. Schuller (2022). "MuSe 2022 Challenge: Multimodal Humour, Emotional Reactions, and Stress". In: *Proc. of the 30th ACM International Conference on Multimedia*. Association for Computing Machinery, pp. 7389–7391 (cit. on pp. 19, 26).

Amiriparian, S., M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller (2017). "Snore Sound Classification Using Image-Based Deep Spectrum Features". In: *Proc. of Interspeech*. ISCA, pp. 3512–3516 (cit. on pp. 12, 30).

Amiriparian, S., A. Sokolov, I. Aslan, L. Christ, M. Gerczuk, T. Hübner, D. Lamanov, M. Milling, S. Ottl, I. Poduremennykh, and B. Schuller, (2021). "On the Impact of Word Error Rate on Acoustic-Linguistic Speech Emotion Recognition: An Update for the Deep Learning Era". in: arXiv:2104.10121 (cit. on pp. 24, 41).

Baevski, A. et al. (2020). "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 12449–12460 (cit. on pp. 2, 13, 22, 31, 40, 50, 52, 60, 62, 74).

Bahdanau, D., (2014). "Neural machine translation by jointly learning to align and translate". in: arXiv preprint arXiv:1409.0473 (cit. on p. 12).

Baird, A., P. Tzirakis, G. Gidel, M. Jiralerspong, E. B. Muller, K. Mathewson, B. Schuller, E. Cambria, D. Keltner, and A. Cowen (2022). *The ICML 2022 Expressive Vocalizations Workshop and Competition: Recognizing, Generating, and Personalizing Vocal Bursts* (cit. on pp. 19, 28, 30).

Baker, K. K., L. O. Ramig, E. S. Luschei, and M. E. Smith, (1998). "Thyroarytenoid muscle activity associated with hypophonia in Parkinson disease and aging". in: Neurology 51(6), pp. 1592–1598 (cit. on p. 59).

Bhati, S., L. M. Velazquez, J. Villalba, and N. Dehak (2019). "LSTM siamese network for Parkinson's disease detection from speech". In: *IEEE global conference on signal and information processing*. IEEE, pp. 1–5 (cit. on p. 59).

Bhattacharjee, T., J. Mallela, Y. Belur, A. Nalini, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh (2021). "Source and Vocal Tract Cues for Speech-Based Classification of Patients with Parkinson's Disease and Healthy Subjects." In: *Proc. of Interspeech*. ISCA (cit. on p. 59).

Bhattacharyya, A., (1943). "On a measure of divergence between two statistical populations defined by their probability distribution". in: Bulletin of the Calcutta Mathematical Society 35, pp. 99–110 (cit. on p. 56).

Bocklet, T., S. Steidl, E. Nöth, and S. Skodda (2013). "Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues." In: *Proc. of Interspeech*. ISCA, pp. 1149–1153 (cit. on p. 59).

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, (2021). "On the opportunities and risks of foundation models". in: arXiv preprint arXiv:2108.07258 (cit. on pp. 13, 49).

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152 (cit. on p. 11).

Brederoo, S., F. Nadema, F. Goedhart, A. Voppel, J. De Boer, J. Wouts, S. Koops, and I. Sommer, (2021). "Implementation of automatic speech analysis for early detection of psychiatric symptoms: what do patients want?" In: Journal of psychiatric research 142, pp. 299–301 (cit. on p. 5).

Breiman, L., (2001). "Random forests". in: Machine learning 45(1), pp. 5–32 (cit. on p. 11).

Burdick, B., C. Holmes, and R. Waln, (1983). "Recognition of suicide signs by physicians in different areas of specialization". in: Journal Of Medical Education (cit. on p. 69).

Burkhardt, F., A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss (2005). “A database of German emotional speech”. In: *Proc. of Interspeech*. ISCA, pp. 1517–1520 (cit. on p. 39).

Busso, C., M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, (2008). “IEMOCAP: Interactive emotional dyadic motion capture database”. in: Language resources and evaluation (cit. on pp. 8, 21, 27, 39, 51).

Busso, C., S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, (2017). “MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception”. in: Transactions on Affective Computing (cit. on pp. 39, 42, 51, 54).

Carletta, J., (2007). “Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus”. in: Language Resources and Evaluation 41, pp. 181–190 (cit. on p. 40).

Chen, S., C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., (2022). “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing”. in: IEEE Journal of Selected Topics in Signal Processing 16(6), pp. 1505–1518 (cit. on pp. 2, 14, 27, 31, 37, 40, 42, 44 sq., 50).

Choi, K., A. Pasad, T. Nakamura, S. Fukayama, K. Livescu, and S. Watanabe (2024). “Self-Supervised Speech Representations are More Phonetic than Semantic”. In: *Proc. of Interspeech*. ISCA, pp. 4578–4582 (cit. on p. 48).

Chopra, S., R. Hadsell, and Y. LeCun (2005). “Learning a similarity metric discriminatively, with application to face verification”. In: *IEEE computer society conference on computer vision and pattern recognition (CVPR)*. Vol. 1. IEEE, pp. 539–546 (cit. on p. 49).

Chowdhury, S. A., N. Durrani, and A. Ali, (2024). “What do end-to-end speech models learn about speaker, language and channel information? a layer-wise and neuron-level analysis”. in: Computer Speech & Language 83, pp. 101539 (cit. on p. 60).

Conneau, A., A. Baevski, R. Collobert, A. Mohamed, and M. Auli (2021). “Unsupervised Cross-Lingual Representation Learning for Speech Recognition”. In: *Proc. of Interspeech*. ISCA, pp. 2426–2430 (cit. on pp. 2, 63, 74).

Cortes, C. and V. Vapnik, (1995). “Support-vector networks”. in: Machine learning 20(3), pp. 273–297 (cit. on p. 11).

Cowen, A., A. Baird, P. Tzirakis, M. Opara, L. Kim, J. Brooks, and J. Metrick (2022). *The Hume Vocal Burst Competition Dataset (H-VB) | Raw Data [A-VB: updated 03.01.22]*. Zenodo (cit. on p. 28).

Cummins, N., S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. Quatieri, (2015). “A review of depression and suicide risk assessment using speech analysis”. in: Speech Communication (cit. on p. 69).

Cummins, N., A. Baird, and B. Schuller, (2018). “Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning”. in: Methods 151, pp. 41–54 (cit. on p. 59).

Cummins, N., J. Epps, M. Breakspear, and R. Goecke (2011). “An investigation of depressed speech detection: Features and normalization”. In: *Proc. of Interspeech*. ISCA (cit. on p. 69).

Darby, J., N. Simmons, and P. Berger, (1984). “Speech and voice parameters of depression: A pilot study”. in: Journal Of Communication Disorders (cit. on p. 78).

Dehak, N., P. J. Kenny, R. OPT, P. Dumouchel, and P. Ouellet, (2010). "Front-end factor analysis for speaker verification". in: IEEE Transactions on Audio, Speech, and Language Processing 19(4), pp. 788–798 (cit. on p. 60).

Dhamyal, H., S. A. Memon, B. Raj, and R. Singh (2020). "The phonetic bases of vocal expressed emotion: natural versus acted". In: *Proc. of Interspeech*. ISCA (cit. on pp. 19, 38).

Drews, M. and M. Krohn (2007). *Plutchik's Wheel of Emotions (Poster)*. Accessed on 2 November 2025. URL: http://www.adliterate.com/archives/Plutchik.emotion.theorie.POSTER.pdf (visited on 11/09/2025) (cit. on p. 7).

Dubagunta, S. P., B. Vlasenko, and M. Magimai.-Doss (2019). "Learning voice source related information for depression detection". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6525–6529 (cit. on pp. 20, 23, 33, 69).

Dušek, P., V. L. y. L. Ibarburu, O. Bezdicek, I. Dall'antonia, S. Dostalova, P. Kovalska, R. Krupička, J. Nepožitek, T. Nikolai, M. Novotnỳ, et al., (2019). "Relations of non-motor symptoms and dopamine transporter binding in REM sleep behavior disorder". in: Scientific reports 9(1), pp. 15463 (cit. on p. 70).

Ekman, P. (1971). "Universals and cultural differences in facial expressions of emotion." In: *Nebraska symposium on motivation*. University of Nebraska Press (cit. on p. 7).

Ekman, P., T. Dalgleish, and M. Power, (1999). "Basic emotions". in: San Francisco, USA (cit. on p. 7).

Elman, J. L., (1990). "Finding structure in time". in: Cognitive science 14(2), pp. 179–211 (cit. on p. 11).

Escobar-Grisales, D., C. D. Ríos-Urrego, I. Baumann, K. Riedhammer, E. Noeth, T. Bocklet, A. M. Garcia, and J. R. Orozco-Arroyave (2024). "It's Time to Take Action: Acoustic Modeling of Motor Verbs to Detect Parkinson's Disease". In: *Proc. of Interspeech*. ISCA, pp. 1965–1969 (cit. on p. 60).

Eyben, F. (2015). "Acoustic features and modelling". In: *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer, pp. 9–122 (cit. on pp. 9, 70, 78).

Eyben, F., K. R. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, (2016). "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing". in: IEEE Transactions on Affective Computing 7(2), pp. 190–202 (cit. on pp. 10, 22, 26, 41, 74).

Eyben, F., M. Wöllmer, and B. Schuller (2010). "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor". In: *Proc. of ACM Multimedia*, pp. 1459–1462 (cit. on pp. 22, 34, 62, 74).

Favaro, A., Y.-T. Tsai, A. Butala, T. Thebaud, J. Villalba, N. Dehak, and L. Moro-Velázquez, (2023). "Interpretable speech features vs. DNN embeddings: What to use in the automatic assessment of Parkinson's disease in multi-lingual scenarios". in: Computers in Biology and Medicine 166, pp. 107559 (cit. on pp. 2, 70, 79).

Feng, K. and T. Chaspari (2024). "A Pilot Study on Clinician-AI Collaboration in Diagnosing Depression from Speech". In: *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, pp. 1–8 (cit. on p. 5).

Feng, T. and S. Narayanan (2023). "Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models". In: *International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 1–8 (cit. on p. 60).

France, D. J., R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes, (2000). "Acoustical properties of speech as indicators of depression and suicidal risk". in: IEEE Transactions On Biomedical Engineering (cit. on p. 69).

French, R. M., (1999). "Catastrophic forgetting in connectionist networks". in: Trends in cognitive sciences 3(4), pp. 128–135 (cit. on p. 49).

Garcia, N., J. R. Orozco-Arroyave, D. Luis Fernando, N. Dehak, and E. Nöth (2017). "Evaluation of the Neurological State of People with Parkinson's Disease Using i-Vectors." In: *Proc. of Interspeech*. ISCA, pp. 299–303 (cit. on p. 59).

Garofolo, J. S., L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium (LDC93S1). DARPA TIMIT corpus of read American English speech (cit. on p. 41).

Gemmeke, J. F., D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter (2017). "Audio set: An ontology and human-labeled dataset for audio events". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 776–780 (cit. on p. 27).

Ghosh, S., E. Laksana, L.-P. Morency, and S. Scherer (2016). "Representation Learning for Speech Emotion Recognition." In: *Proc. of Interspeech*. ISCA, pp. 3603–3607 (cit. on pp. 19, 21).

Goetz, C., B. Tilley, S. Shaftman, G. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. Stern, R. Dodel, et al., (2008). "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results". in: Movement Disorders: Official Journal Of The Movement Disorder Society (cit. on p. 71).

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press (cit. on p. 11).

Grandjean, D., D. Sander, and K. R. Scherer, (2008). "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization". in: Consciousness and cognition 17(2), pp. 484–495 (cit. on p. 7).

Gratch, J., R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al. (2014). "The distress analysis interview corpus of human and computer interviews". In: *LREC* (cit. on p. 71).

Graves, A., A.-r. Mohamed, and G. Hinton (2013). "Speech recognition with deep recurrent neural networks". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 6645–6649 (cit. on p. 11).

Grimm, M., E. Mower, K. Kroschel, and S. Narayanan (2006). "Combining categorical and primitives-based emotion recognition". In: *2006 14th European Signal Processing Conference*. IEEE, pp. 1–5 (cit. on p. 7).

Gunes, H. and B. Schuller, (2013). “Categorical and dimensional affect analysis in continuous input: Current trends and future directions”. in: Image and Vision Computing 31(2), pp. 120–136 (cit. on p. 7).

Gunes, H., B. Schuller, M. Pantic, and R. Cowie (2011). “Emotion representation, analysis and synthesis in continuous space: A survey”. In: *2011 IEEE international conference on automatic face & gesture recognition (FG)*. IEEE, pp. 827–834 (cit. on p. 7).

Gupta, R., T. Chaspari, J. Kim, N. Kumar, D. Bone, and S. Narayanan (2016). “Pathological speech processing: State-of-the-art, current challenges, and future directions”. In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6470–6474 (cit. on p. 59).

Hauptman, Y., R. Aloni-Lavi, I. Lapidot, T. Gurevich, Y. Manor, S. Naor, N. Diamant, and I. Opher (2019). “Identifying Distinctive Acoustic and Spectral Features in Parkinson’s Disease”. In: *Proc. of Interspeech*. ISCA, pp. 2498–2502 (cit. on p. 79).

Hawi, S., J. Alhozami, R. AlQahtani, D. AlSafran, M. Alqarni, and L. El Sahmarany, (2022). “Automatic Parkinson’s disease detection based on the combination of long-term acoustic features and Mel frequency cepstral coefficients (MFCC)”. in: Biomedical Signal Processing and Control 78, pp. 104013 (cit. on p. 59).

He, K., X. Zhang, S. Ren, and J. Sun (2016). “Deep residual learning for image recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (cit. on p. 60).

He, L. and C. Cao, (2018). “Automated depression analysis using convolutional neural networks from speech”. in: J. Biomed. Inform. (cit. on p. 69).

Hinton, G., L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., (2012). “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”. in: IEEE Signal processing magazine 29(6), pp. 82–97 (cit. on p. 1).

Hochreiter, S. and J. Schmidhuber, (1997). “Long short-term memory”. in: Neural Computation 9(8), pp. 1735–1780 (cit. on p. 11).

Hollien, H., (1980). “Vocal indicators of psychological stress”. in: Annals Of The New York Academy Of Sciences (cit. on p. 78).

Hornykiewicz, O., (1998). “Biochemical aspects of Parkinson’s disease”. in: Neurology 51(2_suppl_2), pp. S2–S9 (cit. on p. 59).

Hsu, C.-W. and C.-J. Lin, (2002). “A comparison of methods for multiclass support vector machines”. in: IEEE transactions on Neural Networks 13(2), pp. 415–425 (cit. on p. 11).

Hsu, W.-N., B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, (2021). “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units”. in: IEEE/ACM transactions on audio, speech, and language processing 29, pp. 3451–3460 (cit. on pp. 2, 14, 27, 31, 50).

Hu, E. J., yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen (2022). “LoRA: Low-Rank Adaptation of Large Language Models”. In: *International Conference on Learning Representations* (cit. on pp. 60 sq.).

Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger (2017). "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708 (cit. on p. 27).

Huang, Z., M. Dong, Q. Mao, and Y. Zhan (2014). "Speech emotion recognition using CNN". In: *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 801–804 (cit. on p. 12).

Huggingface (2023a). *w2v2-libri-10min*. `https://huggingface.co/Xinnian/w2v2-libri-10min` (cit. on p. 52).

Huggingface (2023b). *wav2vec2-base-100h*. `https://huggingface.co/facebook/wav2vec2-base-100h` (cit. on p. 52).

Huggingface (2023c). *wav2vec2-base-960h*. `https://huggingface.co/facebook/wav2vec2-base-960h` (cit. on p. 53).

Hussain, M., P. Kumar, S. Khan, D. K. Gordon, and S. Khan, (2020). "Similarities between depression and neurodegenerative diseases: pathophysiology, challenges in diagnosis and treatment options". in: Cureus (cit. on p. 70).

Hussenbocus, A. Y., M. Lech, and N. B. Allen (2015). "Statistical differences in speech acoustics of major depressed and non-depressed adolescents". In: *2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS)*. IEEE, pp. 1–7 (cit. on p. 79).

Janbakhshi, P. (2022). "Automatic pathological speech assessment". Doctoral dissertation. École polytechnique fédérale de Lausanne (cit. on p. 9).

Janbakhshi, P., I. Kodrasi, and H. Bourlard (2021). "Automatic dysarthric speech detection exploiting pairwise distance-based convolutional neural networks". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7328–7332 (cit. on pp. 59, 64).

Joshy, A. A. and R. Rajan (2021). "Automated dysarthria severity classification using deep learning frameworks". In: *European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 116–120 (cit. on p. 59).

Kabil, S. H., H. Muckenhirn, and M. Magimai-Doss (2018). "On Learning to Identify Genders from Raw Speech Signal Using CNNs". In: *Proc. of Interspeech*. ISCA, pp. 287–291 (cit. on pp. 20, 23).

Karaman, O., H. Çakın, A. Alhudhaif, and K. Polat, (2021). "Robust automated Parkinson disease detection based on voice signals with transfer learning". in: Expert Systems with Applications 178, pp. 115013 (cit. on p. 60).

Kearns, J. (2014). *Librivox: Free public domain audiobooks* (cit. on p. 14).

Keesing, A., Y. S. Koh, and M. Witbrock (2021). "Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech." In: *Proc. of Interspeech*. ISCA, pp. 3415–3419 (cit. on p. 42).

Kim, E. and J. W. Shin (2019). "DNN-based emotion recognition based on bottleneck acoustic features and lexical features". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6720–6724 (cit. on p. 19).

Kingma, D. P. and J. L. Ba (2015). "Adam: A method for stochastic gradient descent". In: *ICLR: international conference on learning representations*, pp. 1–15 (cit. on p. 29).

Kirschbaum, C., K.-M. Pirke, and D. H. Hellhammer, (1993). “The ‘Trier Social Stress Test’–a tool for investigating psychobiological stress responses in a laboratory setting”. in: Neuropsychobiology 28(1-2), pp. 76–81 (cit. on p. 26).

Kodrasi, I., (2021). “Temporal envelope and fine structure cues for dysarthric speech detection using CNNs”. in: IEEE Signal Processing Letters 28, pp. 1853–1857 (cit. on p. 59).

Koolagudi, S. and K. Rao, (2012). “Emotion recognition from speech: a review”. in: International journal of speech technology 15(2), pp. 99–117 (cit. on p. 37).

Kornblith, S., J. Shlens, and Q. V. Le (2019). “Do better imagenet models transfer better?” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671 (cit. on p. 49).

Kothare, H., V. Ramanarayanan, M. Neumann, J. Liscombe, V. Richter, L. Lampinen, A. Bai, C. Preciado, K. Brogan, and C. Demopoulos, (2025). “Vocal and facial behavior during affect production in autism spectrum disorder”. in: Journal of Speech, Language, and Hearing Research 68(2), pp. 419–434 (cit. on p. 5).

Kroenke, K., T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, (2009). “The PHQ-8 as a measure of current depression in the general population”. in: Journal of affective disorders (cit. on p. 71).

Kumar, A., A. Raghunathan, R. Jones, T. Ma, and P. Liang, (2022). “Fine-tuning can distort pretrained features and underperform out-of-distribution”. in: arXiv preprint arXiv:2202.10054 (cit. on p. 49).

Kumawat, P. and A. Routray (2021). “Applying TDNN Architectures for Analyzing Duration Dependencies on Speech Emotion Recognition”. In: *Proc. of Interspeech*. ISCA, pp. 3410–3414 (cit. on pp. 19, 24).

La Quatra, M., M. F. Turco, T. Svendsen, G. Salvi, J. R. Orozco-Arroyave, and S. M. Siniscalchi (2024). “Exploiting Foundation Models and Speech Enhancement for Parkinson’s Disease Detection from Speech in Real-World Operative Conditions”. In: *Proc. of Interspeech*. ISCA, pp. 1405–1409 (cit. on p. 65).

Landau, M., (2008). “Acoustical Properties of Speech as Indicators of Depression and Suicidal Risk”. in: Vanderbilt Undergraduate Research Journal (cit. on p. 69).

LeCun, Y. and Y. Bengio (1998). “Convolutional networks for images, speech, and time series”. In: *The Handbook of Brain Theory and Neural Networks*. Ed. by M. A. Arbib. Cambridge, MA, USA: MIT Press, pp. 255–258 (cit. on p. 12).

LeCun, Y., Y. Bengio, and G. Hinton, (2015). “Deep learning”. in: Nature 521(7553), pp. 436–444 (cit. on pp. 1, 59).

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, (1989). “Backpropagation applied to handwritten zip code recognition”. in: Neural Computation 1(4), pp. 541–551 (cit. on p. 12).

Lee, C. M., S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan (2004). “Emotion recognition based on phoneme classes”. In: *Proc. of Interspeech*. ISCA, pp. 889–892 (cit. on pp. 19, 37).

Lee, H. and C. Lyketsos, (2003). “Depression in Alzheimer’s disease: heterogeneity and related issues”. in: Biological Psychiatry (cit. on p. 70).

Lee, K., K. Lee, J. Shin, and H. Lee (2019). "Overcoming Catastrophic Forgetting With Unlabeled Data in the Wild". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 312–321 (cit. on p. 49).

Li, J.-L., T.-Y. Huang, C.-M. Chang, and C.-C. Lee, (2020). "A Waveform-Feature Dual Branch Acoustic Embedding Network for Emotion Recognition". in: Frontiers in Computer Science 2 (cit. on pp. 19, 24).

Li, Y., Y. Mohamied, P. Bell, and C. Lai (2023). "Exploration of a self-supervised speech model: A study on emotional corpora". In: *IEEE Spoken Language Technology Workshop (SLT)*, pp. 868–875 (cit. on pp. 49, 54).

Lim, H., M. J. Kim, and H. Kim (2015). "Robust sound event classification using LBP-HOG based bag-of-audio-words feature representation". In: *Proc. of Interspeech*. ISCA (cit. on p. 10).

Liu, W., Y. Qin, Z. Peng, and T. Lee (2024). "Sparsely Shared Lora on Whisper for Child Speech Recognition". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 11751–11755 (cit. on p. 60).

Logan, B. et al. (2000). "Mel frequency cepstral coefficients for music modeling." In: *Ismir*. Vol. 270. Plymouth, MA, pp. 1–11 (cit. on p. 10).

Lotfian, R. and C. Busso, (2017). "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings". in: IEEE Transactions on Affective Computing 10(4), pp. 471–483 (cit. on p. 8).

Ma, X., H. Yang, Q. Chen, D. Huang, and Y. Wang (2016). "DepAudioNet: An efficient deep model for audio based depression classification". In: *Proceedings of the 6th international workshop on audio/visual emotion challenge (AVEC), On ACM MM* (cit. on p. 69).

Marsella, S. and J. Gratch, (2014). "Computationally modeling human emotion". in: Communications of the ACM 57(12), pp. 56–67 (cit. on p. 7).

McCloskey, M. and N. J. Cohen (1989). "Catastrophic interference in connectionist networks: The sequential learning problem". In: *Psychology of learning and motivation*. Vol. 24. Elsevier, pp. 109–165 (cit. on p. 49).

Mehrabian, A., (1996). "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament". in: Current psychology 14(4), pp. 261–292 (cit. on p. 7).

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR* (cit. on p. 38).

Mirsamadi, S., E. Barsoum, and C. Zhang (2017). "Automatic speech emotion recognition using recurrent neural networks with local attention". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2227–2231 (cit. on p. 12).

Mohamed, A., H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, et al., (2022). "Self-supervised speech representation learning: A review". in: IEEE Journal of Selected Topics in Signal Processing (cit. on pp. 1, 13, 49, 60).

Moore II, E., M. A. Clements, J. W. Peifer, and L. Weisser, (2007). “Critical analysis of the impact of glottal features in the classification of clinical depression in speech”. in: IEEE Transactions On Biomedical Engineering (cit. on p. 69).

Moro-Velazquez, L., J. Villalba, and N. Dehak (2020). “Using x-vectors to automatically detect parkinson’s disease from speech”. In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1155–1159 (cit. on p. 59).

Muckenhirn, H., M. Magimai.-Doss, and S. Marcel (2018a). “Towards directly modeling raw speech signal for speaker verification using CNNs”. In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4884–4888 (cit. on pp. 20, 23, 32 sq.).

Muckenhirn, H., V. Abrol, M. M. Doss, and S. Marcel (2018b). *Gradient-based spectral visualization of CNNs using raw waveforms*. Tech. rep. Idiap (cit. on p. 33).

Muckenhirn, H., V. Abrol, M. Magimai-Doss, and S. Marcel (2019). “Understanding and Visualizing Raw Waveform-based CNNs”. In: *Proc. of Interspeech*. ISCA (cit. on p. 33).

Mukhoti, J., Y. Gal, P. H. Torr, and P. K. Dokania, (2023). “Fine-tuning can cripple your foundation model; preserving features may be the solution”. in: arXiv preprint arXiv:2308.13320 (cit. on p. 49).

Neumann, M. and N. T. Vu (2019a). “Improving speech emotion recognition with unsupervised representation learning on unlabeled speech”. In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7390–7394 (cit. on p. 21).

Neumann, M. and T. Vu (2017). “Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech”. In: *Proc. of Interspeech*. ISCA, pp. 1263–1267 (cit. on pp. 19, 24, 37).

Neumann, M. and N. T. Vu (2019b). “Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech”. In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7390–7394 (cit. on pp. 42, 54).

Ngo, Q. C. et al., (2022). “Computerized analysis of speech and voice for Parkinson’s disease: A systematic review”. in: Computer Methods and Programs in Biomedicine 226, pp. 107133 (cit. on p. 59).

Nickerson, C. A., (1997). “A note on" A concordance correlation coefficient to evaluate reproducibility"”. in: Biometrics, pp. 1503–1507 (cit. on p. 15).

Nilsonne, Å., J. Sundberg, S. Ternström, and A. Askenfelt, (1988). “Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression”. in: The Journal of the Acoustical Society of America (cit. on p. 69).

Orozco-Arroyave, J. R., J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth (2014). “New Spanish speech corpus database for the analysis of people suffering from Parkinson’s disease.” In: *Lrec* (cit. on p. 62).

Orozco-Arroyave, J. R., J. C. Vásquez-Correa, J. F. Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei, et al., (2018). “NeuroSpeech: An open-source software for Parkinson’s speech analysis”. in: Digital Signal Processing 77, pp. 207–221 (cit. on p. 74).

Ozkanca, Y., M. Göksu Öztürk, M. Ekmekci, D. Atkins, C. Demiroglu, and R. Hosseini Ghomi, (2019). "Depression screening from voice samples of patients affected by parkinson's disease". in: Digital Biomarkers (cit. on p. 70).

Palaz, D., M. Magimai.-Doss, and R. Collobert, (2019). "End-to-End Acoustic Modeling using Convolutional Neural Networks for HMM-based Automatic Speech Recognition". in: Speech Communication 108, pp. 15–32 (cit. on pp. 1, 20, 23, 32 sq., 35).

Panayotov, V., G. Chen, D. Povey, and S. Khudanpur (2015). "Librispeech: An ASR corpus based on public domain audio books". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5206–5210 (cit. on pp. 14, 41, 52).

Pasad, A., C.-M. Chien, S. Settle, and K. Livescu, (2024). "What do self-supervised speech models know about words?" In: Transactions of the Association for Computational Linguistics 12, pp. 372–391 (cit. on p. 50).

Pasad, A., J.-C. Chou, and K. Livescu (2021). "Layer-Wise Analysis of a Self-Supervised Speech Representation Model". In: *Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 914–921 (cit. on p. 50).

Pasad, A., B. Shi, and K. Livescu (2023). "Comparative layer-wise analysis of self-supervised speech models". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE (cit. on p. 60).

Peng, Z. et al. (2021). "Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3020–3024 (cit. on pp. 19, 37).

Pepino, L., P. Riera, and L. Ferrer (2021). "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings". In: *Proc. of Interspeech*. ISCA, pp. 3400–3404 (cit. on pp. 37, 42, 49 sq., 54).

Pérez-Toro, P., J. Vásquez-Correa, T. Bocklet, E. Nöth, and J. Orozco-Arroyave, (2021). "User state modeling based on the arousal-valence plane: Applications in customer satisfaction and health-care". in: IEEE Transactions On Affective Computing (cit. on pp. 70 sq., 74 sqq.).

Pérez-Toro, P. A., T. Arias-Vergara, P. Klumpp, J. C. Vásquez-Correa, M. Schuster, E. Noeth, and J. R. Orozco-Arroyave, (2022). "Depression assessment in people with Parkinson's disease: The combination of acoustic features and natural language processing". in: Speech Communication 145, pp. 10–20 (cit. on p. 70).

Pérez-Toro, P. A., D. Rodríguez-Salas, T. Arias-Vergara, S. P. Bayerl, P. Klumpp, K. Riedhammer, M. Schuster, E. Nöth, A. Maier, and J. R. Orozco-Arroyave (2023). "Transferring quantified emotion knowledge for the detection of depression in Alzheimer's disease using forestnets". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on p. 70).

Plutchik, R. (1982). *A psychoevolutionary theory of emotions* (cit. on p. 7).

Plutchik, R., (2001). "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice". in: American scientist 89(4), pp. 344–350 (cit. on p. 7).

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution*. American Psychological Association (cit. on p. 7).

Prabhakera, N. N. and P. Alku (2018). “Dysarthric speech classification using glottal features computed from non-words, words and sentences”. In: *Proc. of Interspeech*. ISCA, pp. 3403–3407 (cit. on p. 59).

Purohit, T., I. Ben Mahmoud, B. Vlasenko, and M. Magimai-Doss (2022). “Comparing supervised and self-supervised embedding for ExVo Multi-Task learning track”. In: *Proceedings of the ICML Expressive Vocalizations Workshop held in conjunction with the 39th International Conference on Machine Learning*. Maryland, USA (cit. on p. 3).

Purohit, T. and M. Magimai-Doss (2025). “Emotion information recovery potential of wav2vec2 network fine-tuned for speech recognition task”. In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on p. 4).

Purohit, T., B. Ruvolo, J. R. Orozco-Arroyave, and M. M. Doss (2025a). “Automatic Parkinson’s disease detection from speech: Layer selection vs adaptation of foundation models”. In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on pp. 4, 79).

Purohit, T., B. Ruvolo, J. R. Orozco-Arroyave, and M. Magimai-Doss (2025b). “On Detection of Depression in Parkinson’s Disease Patients’ Speech: Handcrafted Features vs. Speech Foundation Models”. In: *Automatic Assessment of Parkinsonian Speech*. Ed. by J. I. G. Llorente. 1st ed. Vol. 2646. Communications in Computer and Information Science (CCIS). Cambridge, MA, USA: Springer Nature Switzerland AG (cit. on p. 4).

Purohit, T., B. Vlasenko, and M. Magimai.-Doss (2023a). “Implicit phonetic information modeling for speech emotion recognition”. In: *Proc. of Interspeech*. ISCA, pp. 1883–1887 (cit. on pp. 3, 53 sq.).

Purohit, T., S. Yadav, B. Vlasenko, S. P. Dubagunta, and M. Magimai.-Doss (2023b). “Towards Learning Emotion Information from Short Segments of Speech”. In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on pp. 3, 41, 54).

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever (2023). “Robust speech recognition via large-scale weak supervision”. In: *International conference on machine learning*. PMLR, pp. 28492–28518 (cit. on p. 63).

Ramet, G., P. N. Garner, M. Baeriswyl, and A. Lazaridis (2018). “Context-aware attention mechanism for speech emotion recognition”. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 126–131 (cit. on p. 13).

Ringeval, F., F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, (2015a). “Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data”. in: Pattern Recognition Letters 66, pp. 22–30 (cit. on p. 15).

Ringeval, F., B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic (2015b). “Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data”. In: *Proceedings of the 5th international workshop on audio/visual emotion challenge*, pp. 3–8 (cit. on p. 15).

Rios-Urrego, C. D., S. A. Moreno-Acevedo, E. Nöth, and J. R. Orozco-Arroyave (2022). “End-to-end Parkinson’s disease detection using a deep convolutional recurrent network”. In: *International Conference on Text, Speech, and Dialogue*. Springer, pp. 326–338 (cit. on p. 59).

Rodríguez-Salas, D., N. Mürschberger, N. Ravikumar, M. Seuret, and A. Maier, (2020). "Mapping ensembles of trees to sparse, interpretable multilayer perceptron networks". in: SN Computer Science (cit. on p. 70).

Rosenblatt, F., (1958). "The perceptron: A probabilistic model for information storage and organization in the brain". in: Psychological Review 65(6), pp. 386–408 (cit. on p. 11).

Rozgić, V. et al. (2012). "Ensemble of SVM trees for multimodal emotion recognition". In: *Proceedings Asia Pacific Signal and Information processing Association Annual Summit and Conference*. IEEE, pp. 1–4 (cit. on p. 21).

Rumelhart, D. E., G. E. Hinton, and R. J. Williams, (1986). "Learning representations by back-propagating errors". in: nature 323(6088), pp. 533–536 (cit. on p. 11).

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., (2015). "Imagenet large scale visual recognition challenge". in: International journal of computer vision 115(3) (cit. on pp. 27, 30, 60).

Russell, J. A., (1980). "A circumplex model of affect." in: Journal of personality and social psychology 39(6), pp. 1161 (cit. on p. 7).

Russell, J. A. and A. Mehrabian, (1977). "Evidence for a three-factor theory of emotions". in: Journal of research in Personality 11(3), pp. 273–294 (cit. on p. 8).

Rusz, J., R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, (2013). "Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task". in: The Journal of the Acoustical Society of America 134(3), pp. 2171–2181 (cit. on p. 59).

Saeed, A., D. Grangier, and N. Zeghidour (2021). "Contrastive learning of general-purpose audio representations". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 3875–3879 (cit. on p. 27).

Scherer, K. R., (1978). "Personality inference from voice quality: The loud voice of extroversion". in: European Journal of Social Psychology 8(4), pp. 467–487 (cit. on p. 1).

Scherer, K. R., (2003). "Vocal communication of emotion: A review of research paradigms". in: Speech communication 40(1-2), pp. 227–256 (cit. on p. 37).

Scherer, K. R., (2005). "What are emotions? And how can they be measured?" In: Social science information 44(4), pp. 695–729 (cit. on p. 37).

Schmitt, M., F. Ringeval, and B. Schuller (2016). "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech." In: *Proc. of Interspeech*. ISCA, pp. 495–499 (cit. on p. 10).

Schmitt, M. and B. Schuller, (2017). "openXBOW - Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit". in: Journal of Machine Learning Research (cit. on pp. 10, 40, 52).

Schuller, B., (2018). "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends". in: Communications of the ACM 61(5), pp. 90–99 (cit. on pp. 1, 6, 9, 37).

Schuller, B. and A. Batliner (2013). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons (cit. on pp. 1, 5 sq., 9 sq., 19).

Schuller, B., A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson (2007). "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals". In: *Proc. of Interspeech*. ISCA (cit. on p. 37).

Schuller, B., S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, (2013a). "Paralinguistics in speech and language—state-of-the-art and the challenge". in: Computer Speech & Language 27(1), pp. 4–39 (cit. on pp. 1, 5, 8, 37).

Schuller, B., S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang (2014). "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load, multitasking". In: *Proc. of Interspeech*. ISCA (cit. on pp. xi, 34 sq.).

Schuller, B., S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim (2013b). "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism". In: *Proc. of Interspeech*. ISCA (cit. on pp. 10, 22, 30, 40, 52, 74).

Schuller, B., B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth (2009). "Acoustic emotion recognition: A benchmark comparison of performances". In: *Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 552–557 (cit. on pp. 39, 41).

Scibelli, F., G. Roffo, M. Tayarani, L. Bartoli, G. Mattia, A. Esposito, and A. Vinciarelli (2018). "Depression Speaks: Automatic Discrimination between Depressed and Non-Depressed Speakers Based on Nonverbal Speech Features". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE (cit. on p. 69).

Simonyan, K. and A. Zisserman (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, pp. 1–14 (cit. on p. 30).

Snyder, D., D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur (2018). "X-vectors: Robust dnn embeddings for speaker recognition". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5329–5333 (cit. on p. 60).

Song, Z. et al. (2024). "LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR". In: *Proc. of Interspeech*. ISCA, pp. 3934–3938 (cit. on p. 60).

Springenberg, J. T. et al. (2015). "Striving for Simplicity: The All Convolutional Net". In: *ICLR (Workshop)* (cit. on p. 33).

Stappen, L. et al. (2021). "MuSe 2021 Challenge: Multimodal Emotion, Sentiment, Physiological-Emotion, and Stress Detection". In: *Proc. of ACM Multimedia*. Association for Computing Machinery, pp. 5706–5707 (cit. on p. 19).

Stasak, B., J. Epps, N. Cummins, and R. Goecke (2016). "An investigation of emotional speech in depression classification". In: *Proc. of Interspeech*. ISCA (cit. on p. 69).

Steidl, S., A. Batliner, B. Schuller, and D. Seppi (2009). "The hinterland of emotions: facing the open-microphone challenge". In: *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, pp. 1–8 (cit. on p. 8).

Tracy, J. L. and D. Randles, (2011). “Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt”. in: Emotion review 3(4), pp. 397–405 (cit. on p. 7).

Trigeorgis, G., F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou (2016). “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network”. In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5200–5204 (cit. on pp. 11 sq.).

Tzirakis, P., G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, (2017). “End-to-end multimodal emotion recognition using deep neural networks”. in: IEEE Journal of selected topics in signal processing 11(8), pp. 1301–1309 (cit. on p. 11).

Valstar, M., J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic (2016). “Avec 2016: Depression, mood, and emotion recognition workshop and challenge”. In: *AVEC* (cit. on pp. 74, 76).

Vapnik, V. and A. Y. Lerner, (1963). “Recognition of patterns with help of generalized portraits”. in: Avtomat. i Telemekh 24(6), pp. 774–780 (cit. on p. 11).

Vásquez-Correa, J. C., J. R. Orozco-Arroyave, and E. Nöth (2017). “Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson’s Disease.” In: *Proc. of Interspeech*. ISCA, pp. 314–318 (cit. on p. 59).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, (2017). “Attention is all you need”. in: Advances in neural information processing systems 30 (cit. on p. 12).

Virmani, T., M. Lotia, A. Glover, L. Pillai, A. S. Kemp, A. Iyer, P. Farmer, S. Syed, L. J. Larson-Prior, and F. W. Prior, (2022). “Feasibility of telemedicine research visits in people with Parkinson’s disease residing in medically underserved areas”. in: Journal of Clinical and Translational Science 6(1), pp. e133 (cit. on p. 59).

Vlasenko, B. and A. Wendemuth (2013). “Determining the smallest emotional unit for level of arousal classification”. In: *Proc. of ACII*. HUMAINE Association. IEEE, pp. 511–516 (cit. on pp. 19, 37 sq.).

Vlasenko, B., D. Prylipko, R. Böck, and A. Wendemuth, (2014). “Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications”. in: Computer Speech & Language (cit. on p. 38).

Wagner, D., I. Baumann, F. Braun, S. P. Bayerl, E. Nöth, K. Riedhammer, and T. Bocklet (2023). “Multi-class Detection of Pathological Speech with Latent Features: How does it perform on unseen data?” In: *Proc. of Interspeech*. ISCA, pp. 2318–2322 (cit. on pp. 60, 63).

Wiepert, D. A., R. L. Utianski, J. R. Duffy, J. L. Stricker, L. R. Barnard, D. T. Jones, and H. Botha (2024). “Speech foundation models in healthcare: Effect of layer selection on pathological speech feature prediction”. In: *Proc. of Interspeech*. ISCA, pp. 4618–4622 (cit. on p. 60).

Wodzinski, M., A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave, and E. Nöth (2019). “Deep learning approach to Parkinson’s disease detection using voice recordings and convolutional neural network dedicated to image classification”. In: *Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, pp. 717–720 (cit. on p. 60).

World Health Organization (2023). *Depression*. URL: https://www.who.int/news-room/fact-sheets/detail/depression (visited on 10/19/2023) (cit. on p. 69).

World Health Organization (2025). *Global Health Estimates: Life expectancy and leading causes of death and disability*. URL: https://www.who.int/health-topics/mental-health (visited on 01/15/2025) (cit. on p. 69).

Wu, W., C. Zhang, and P. C. Woodland (2023). "Self-supervised representations in speech-based depression detection". In: *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cit. on pp. 50, 69, 79).

Xia, R. and Y. Liu, (2015). "A multi-task learning framework for emotion recognition using 2D continuous space". in: IEEE Transactions on affective computing 8(1), pp. 3–14 (cit. on p. 21).

Xia, R. and Y. Liu (2016). "DBN-ivector Framework for Acoustic Emotion Recognition". In: *Proc. of Interspeech*. ISCA, pp. 480–484 (cit. on pp. 19, 24).

Yadav, S., T. Purohit, Z. Mostaani, B. Vlasenko, and M. Magimai-Doss (2022). "Comparing Biosignal and Acoustic feature Representation for Continuous Emotion Recognition". In: *International Multimodal Sentiment Analysis Workshop and Challenge* (cit. on p. 3).

Yang, S.-w., P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee (2021). "SUPERB: Speech Processing Universal PERformance Benchmark". In: *Proc. of Interspeech*. ISCA, pp. 1194–1198 (cit. on pp. 2, 13, 31, 41 sq., 44, 49, 53).

Yenigalla, P., A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa (2018). "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding". In: *Proc. of Interspeech*. ISCA (cit. on p. 38).

Yu, D., M. L. Seltzer, J. Li, J.-T. Huang, and S. Frank (2013). "Feature Learning in Deep Neural Networks: Studies on Speech Recognition Tasks". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. Scottsdale, Arizona, USA (cit. on p. 12).

Yuan, J., X. Cai, R. Zheng, L. Huang, and K. Church, (2021). "The role of phonetic units in speech emotion recognition". in: arXiv preprint arXiv:2108.01132 (cit. on pp. 19, 38).

Zahid, L., M. Maqsood, S. S. Farooq, F. Aadil, I. Mehmood, M. Fiaz, and S. K. Jung (2020). "Detection of Speech Impairments in Parkinson Disease Using Handcrafted Feature-Based Model on Spanish Speech Corpus". In: *International Workshop on Frontiers of Computer Vision*. Springer, pp. 54–65 (cit. on p. 69).

Zhai, X., J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, et al., (2019). "A large-scale study of representation learning with the visual task adaptation benchmark". in: arXiv preprint arXiv:1910.04867 (cit. on p. 49).

Zhang, Z., W. Xu, Z. Dong, K. Wang, Y. Wu, J. Peng, R. Wang, and D.-Y. Huang, (2024). "ParaL-Bench: A large-scale benchmark for computational paralinguistics over acoustic foundation models". in: IEEE Transactions on Affective Computing (cit. on p. 13).

Zhao, J., X. Mao, and L. Chen, (2019). "Speech emotion recognition using deep 1D & 2D CNN LSTM networks". in: Biomedical Signal Processing and Control 47, pp. 312–323 (cit. on p. 19).

Zheng, H., L. Shen, A. Tang, Y. Luo, H. Hu, B. Du, and D. Tao, (2023). “Learn from model beyond fine-tuning: A survey”. in: arXiv preprint arXiv:2310.08184 (cit. on p. 49).

Tilak Purohit

Idiap Research Institute, Rue Marconi 19, 1920 Martigny, Switzerland
• tilak.purohit@gmail.com • Google Scholar • ORCiD • LinkedIn

EDUCATION

EPFL, the Swiss Federal Institute of Technology in Lausanne* — **Lausanne, Switzerland**
- Doctoral Student - Electrical Engineering — 2021 – 2025

IIITB, International Institute of Information Technology Bangalore — **Bangalore, India**
- Masters in Technology - Computer Science — 2019

EXPERIENCE

Idiap Research Institute — **Martigny, Switzerland**
Research Assistant, at the Speech and Audio Processing Group — 2021– Present
- **Supervisor:** Dr. Mathew Magimai.-Doss
  - Developing speech processing methods to assess non-motor aspects of speech in neurotypical & atypical individuals.
  - Research funded by the Swiss National Science Foundation (SNSF) through the Bridge Discovery project - EMIL: Emotion in the loop - a step towards a comprehensive closed-loop deep brain stimulation in Parkinson's disease.

Amazon AGI Foundations (United States) — **Boston, USA**
Applied Scientist Intern, AGI foundations- Speech & Audio — May 2025 – Aug 2025
- **Mentor:** Dr. Michael Owen **Manager:** Dr. Andrew Fletcher
  - Designed and trained reward models for the Reinforcement Learning with Human Feedback (RLHF) pipeline, and implemented Supervised Fine-Tuning (SFT) baselines to align Amazon's Nova Sonic (a speech-to-speech foundation model for conversational AI) with human feedback to enhance expressivity. Collaborated with multiple teams and stakeholders across the Amazon AGI Speech and Audio organization.

Universidad de Antioquia (UdeA) — **Medellín, Colombia**
Visiting research scholar at GITA Lab, Faculty of Engineering — May 2024 – Jul 2024
- **Hosted by:** Prof. Juan Rafael Orozco Arroyave
  - Investigating the ON/OFF effects of levodopa on the speech of Parkinson's disease patients, analyzing both motor and non-motor speech aspects. Modeling longitudinal data.

Indian Institute of Science (IISc) — **Bangalore, India**
Research Assistant at SPIRE Lab, Electrical Engineering Department — 2019 – 2021
- **Supervisor:** Prof. Prasanta Kumar Ghosh
  - Analysis, visualization, and inversion of speech articulation. Notably, created an acoustic-articulatory corpus.
  - Research funded by Department of Science and Technology (DST), Govt. of India.

DOCTORAL THESIS

Title: *"Advancing Neural Representations for Paralinguistic Analysis: From Speech Emotion to Parkinson's Disease Assessment"*
Supervisor: Dr. Mathew Magimai Doss & Prof. Jean-Philippe Thiran
Jury: Prof. Shrikanth Narayanan, Prof. Mahsa Shoran, Prof. Isabel Trancoso, and Dr. Jean-Marc Odobez.

MASTERS THESIS

Title: *"Temporal Decomposition of Speech - with applications to co-articulation modeling"*
Supervisor: Prof. V Ramasubramanian
Jury: Prof. D.B. Jayagopi & Prof. G Srinivasaraghavan.

COMPUTER SKILLS

Languages: Python
Tools & Frameworks: PyTorch, Scikit-learn, OpenSmile.

AWARDS & ACHIEVEMENTS

- First place in the Lemanic Life Sciences Hackathon 2024, winning CHF 1000 as a cash prize. Project proposal and team lead: **Tilak Purohit**.
- Runner-up team (2^{nd} place) in Innosuisse Start-up training program 2023, a 14-week federal program for start-up founders, organized by EPFL Innovation Park. Proposed Emoscan, a mental health monitoring application. Project proposal and team lead: **Tilak Purohit**.
- Ranked among the top 5 systems submitted at the ICML Expressive Vocalizations (ExVo) Competition, 2022.
- Best paper award recipient at SPCOM 2020. Paper presentation and talk delivered by **Tilak Purohit**.
- All India Council for Technical Education (AICTE), Govt. of India, scholarship holder for 2 years.

*EPFL, l'École polytechnique fédérale de Lausanne

- All India Rank 816 among 120,000 computer science undergraduate candidates in the GATE 2017 exam, a national-level postgraduate engineering admission examination.

TALKS

- Modeling Speech Variability: From Speech Articulatory Dynamics to Paralinguistics — **UCSF, USA**, Aug 2025
  Hosted by: Dr. David Moses and Dr. Jessie Liu, at Chang Lab (On-site)
- Teaching machines to decode human emotions through speech — **UdeA, Colombia**, May 2023
  Hosted by: Prof. Juan Rafael Orozco Arroyave, GITA symposium talk (On-site)

POSTER PRESENTATIONS

Selected venues where research has been featured:

- Valais/Wallis Workshop on Artificial Intelligence, Idiap, Switzerland — Apr 2025
- Applied Machine Learning Days (AMLD), EPFL, Switzerland — Feb 2025
- Evolving Language NCCR Summer School, Grindelwald, Switzerland — May 2023

ACADEMIC MENTORSHIP

- Co-supervised Master's thesis of Barbara Ruvolo on "Interpretable acoustic features for depression detection: a comparative study of healthy & Parkinson's disease individuals," guiding research design, data analysis, and scientific writing (2023–24); mentorship contributed to conference and workshop publications.
- Co-supervised C. Siddarth and Arvind Ramesh at IISc Bangalore (2021); mentorship contributed to conference publications.

TEACHING AND ACADEMIC SERVICE

- Guest Lecture: "Computational Paralinguistics: Methods and Emerging Trends" | EPFL EE-554: Automatic Speech Processing | 2024-2025
- Teaching Assistant (Labs and Evaluation) | EPFL EE-559: Deep Learning | 2024
- Teaching Assistant (Evaluation) | UniDistance MO9: Introduction to Speech Processing | 2022

REVIEWING

IEEE OJSP (2026), ICASSP (2023, 2024, 2025), Interspeech (2023, 2024, 2025), EUSIPCO (2023)

DATASET

SPIRE-VCV: An acoustic-articulatory corpus with three different speaking rates [Available upon request]

- Created an acoustic-articulatory dataset of Vowel Consonant Vowel (VCV) utterances, recorded at 3 different speaking rates. The dataset includes recordings from 10 non-native English speakers (5 female & 5 male) aged 18-22, featuring combinations of 5 vowels and 17 consonants with manual VCV boundary annotations.
- Articulatory data was recorded using a 3D Electromagnetic Articulograph (EMA) AG501.

BOOK CHAPTER

- Tilak Purohit, Barbara Ruvolo, Juan Rafael Orozco-Arroyave, and Mathew Magimai-Doss. "On Detection of Depression in Parkinson's Disease Patients' Speech: Handcrafted Features vs. Speech Foundation Models". In Juan Ignacio Godino Llorente, editor, *Automatic Assessment of Parkinsonian Speech*, pages 103–117, Cham, 2026. Springer Nature Switzerland

CONFERENCE PUBLICATIONS

- Tilak Purohit and Mathew Magimai Doss. "Emotion information recovery potential of wav2vec2 network fine-tuned for speech recognition task". In *Proc. of **ICASSP, 2025**, Hyderabad, India.*
- Tilak Purohit, Barbara Ruvolo, Juan Rafael Orozco-Arroyave, and Mathew Magimai Doss. "Automatic Parkinson's disease detection from speech: Layer selection vs adaptation of foundation models". In *Proc. of **ICASSP, 2025**, Hyderabad, India.*
- Luis Parra Gallego, Tilak Purohit, Bogdan Vlasenko, Mathew Magimai Doss, and Juan Rafael Orozco-Arroyave. "Cross-transfer Knowledge Between Speech and Text Encoders to Evaluate Customer Satisfaction". In *Proc. of **Interspeech, 2024**, Kos Island, Greece.*
- Tilak Purohit, Bogdan Vlasenko, and Mathew Magimai Doss. "Implicit phonetic information modeling for speech emotion recognition". In *Proc. of **Interspeech, 2023**, Dublin, Ireland.*
- C Siddarth, Arvind Ramesh, Tilak Purohit, and Prasanta Kumar Ghosh. "A Study on the Importance of Formant Transitions for Stop-Consonant Classification in VCV Sequence". In *Proc. of **Interspeech, 2023**, Dublin, Ireland.*
- Tilak Purohit, Sarthak Yadav, Bogdan Vlasenko, S Pavankumar Dubagunta, and Mathew Magimai Doss. "Towards Learning Emotion Information from Short Segments of Speech". In *Proc. of **ICASSP, 2023**, Rhodes island, Greece.*
- Tilak Purohit, Tejas Umesh, Shankar Narayanan, S Minulakshmi, and Prasanta Kumar Ghosh. "SPIRE VCV: An Acoustic-Articulatory Corpus with Three Different Speaking Rates". In *Proc. of **24th O-COCOSDA, 2021**, NUS, Singapore.* (SPIRE VCV Corpus release)

- Tilak Purohit, Achuth Rao MV, and Prasanta Kumar Ghosh. "Impact of speaking rate on the source filter interaction in speech: A Study". In *Proc. of **ICASSP, 2021**, Toronto, Canada.*
- Tilak Purohit and Prasanta Kumar Ghosh. "An Investigation of the Virtual Lip Trajectories During the Production of Bilabial Stops and Nasal at Different Speaking Rates". In *Proc. of **Interspeech, 2020**, Shanghai, China.*
- Tilak Purohit and V. Ramasubramanian. "Component-specific temporal decomposition: application to enhanced speech coding and co-articulation analysis". In *Proc. of **SPCOM, 2020**, IISc, Bangalore, India.*
- T Kumar, S Sundar, Tilak Purohit*, and V. Ramasubramanian. "End-to-end audio-scene classification from raw audio: Multi time-frequency resolution CNN architecture for efficient representation learning". In *Proc. of **SPCOM, 2020**, IISc, Bangalore, India.* **(Equal Contribution)** **(Best Paper Award)** (Presented by: Tilak Purohit)

WORKSHOP PUBLICATIONS

- Joanna Reszka, Parvaneh Janbakhshi, Tilak Purohit, and Sadegh Mohammadi. "Investigating the Effects of Diffusion-based Speech Enhancement Models on Dysarthric Speech". In *Proc. of **SPADE workshop of ICASSP, 2025**, Hyderabad, India.* **(Best Paper Award)**
- Barbara Ruvolo, Tilak Purohit, Bogdan Vlasenko, and Mathew Magimai Doss. "Exploring the Complexity of Parkinson Patients' Speech for Depression Detection task: A Qualitative Analysis". In *Proc. of **SPADE workshop of ICASSP, 2025**, Hyderabad, India.*
- Sarthak Yadav, Tilak Purohit, Zohreh Mostaani, Bogdan Vlasenko, and Mathew Magimai Doss. Comparing biosignal and acoustic feature representation for continuous emotion recognition. In *Proc of **ACM International Conference on Multimedia: 3rd MuSe Workshop and Challenge, 2022**, Lisboa, Portugal.*
- Tilak Purohit, Imen Ben Mahmoud, Bogdan Vlasenko, and Mathew Magimai Doss. "Comparing supervised and self-supervised embedding for ExVo Multi-Task learning track". In *Proc of **ICML Expressive Vocalizations (ExVo) Workshop and Competition, 2022**, Baltimore, Maryland, USA.*
- Kaajal Gupta, Tilak Purohit, Anzar Zulfiqar, Pushpa Ramu, and V. Ramasubramanian. "Detection of emotional states of OCD patients in an exposure-response prevention therapy scenario". In *Proc. of - **SMM, Workshop of Interspeech, 2019**, Graz, Austria.*, 2019 (Presented by: Kaajal Gupta & Tilak Purohit)
- Tilak Purohit, Atul Agrawal, and V. Ramasubramanian. "Acoustic scene classification using deep CNN on raw-waveform". *Tech. Rep. **DCASE Workshop and Challenge, 2018**, Surrey, UK.*

[*CV compiled on 2026-01-14*]