Proximal Interpolative Fusion of Heterogeneous Representations for Affective and Secure Voice Intelligence

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

ORCHID CHETIA PHUKAN

PHD22025

COMPUTER SCIENCE AND ENGINEERING

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

June 14, 2026

THESIS CERTIFICATE

This is to certify that the thesis titled **Proximal Interpolative Fusion of Heterogeneous Representations for Affective and Secure Voice Intelligence**, submitted by **Orchid Chetia Phukan**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Arun Balaji Buduru
Supervisor
Associate Professor
Dept. of Computer Science and Engineering
IIIT-Delhi, India

Prof. Rajesh Sharma
Supervisor
Professor
Dept. of Computer Science and Artificial Intelligence
Plaksha University, India

Place: New Delhi

Date: 14th June, 2026

Declaration

"I acknowledge that I am fully responsible for the entire content of my thesis, including any sections assisted by online tools, including Artificial Intelligence-based tools. I accept full accountability for any violations of ethical standards in publications arising from the use of such tools.".

Month, Year: June, 2026

Student Name: Orchid Chetia Phukan 

Roll Number: PhD22025

Advisor(s) Name: Dr. Arun Balaji Buduru & Prof. Rajesh Sharma

Indraprastha Institute of Information Technology Delhi

New Delhi 110020

ACKNOWLEDGEMENTS

Looking back now, this Ph.D. has gone by like the wind—years that passed almost unnoticed because I was so deeply immersed in research, experiments, and ideas. It feels surreal to be writing this acknowledgment, marking the close of a journey that was intense, demanding, and deeply fulfilling. In many ways, this also feels like a moment to thank myself—for the persistence and commitment that carried me through.

I am thankful to everyone who has supported, encouraged, and believed in me throughout this journey. I am thankful to my advisors, Dr. Arun Balaji Buduru and Professor Rajesh Sharma. Their guidance has helped me grow as a researcher and also as a individual. I am especially thankful to Dr. Arun Balaji Buduru for giving me the freedom to explore the problems I genuinely care about—this freedom allowed me to follow my instincts, learn independently, and build work that reflects my passion. Both my advisors consistently pushed me intellectually and supported me with patience and unwavering belief.

I am thankful to my institute, IIIT-Delhi, for their generous support through the stipend and conference funding throughout the years. Their financial assistance removed barriers and enabled me to pursue research without distraction. Their financial help took away obstacles and let me do research without being distracted. I also appreciate the chances I've had to present my works at international conferences like INTERSPEECH 2023, INTERSPEECH 2024, EUSIPCO 2024, INTERSPEECH 2025, ICASSP 2025, and APSIPA 2025. The welcoming speech community has been a big part of my academic path and I look forward to contributing more to this field in the years ahead. I'm also grateful for the travel grants that supported me for presenting my research at INTERSPEECH 2024, EUSIPCO 2024, and APSIPA 2025.

Finally, I want to express my deepest gratitude to my parents—Maa and Deuta—my younger brother, my closest research comrades Mohd Mujtaba Akhtar and Girish, and my lab-mates in USG. Their constant support, encouragement, and belief in me kept me grounded, especially during the most demanding moments. This journey would

not have been possible without their presence, patience, and faith. I would also like to express my sincere thanks to my defense committee, Prof. Hemant Purohit, Prof. Salil Kanhere, and Prof. Anil Kumar Vuppala for providing valuable feedback and suggestions.

Orchid Chetia Phukan

ABSTRACT

KEYWORDS: Voice Intelligence; Pre-Trained Models; Emotion Recognition; Synthetic Speech Detection

The rapid emergence of large-scale Pre-Trained Models (PTMs) trained on massive and diverse corpora has fundamentally reshaped the landscape of numerous AI domains, including natural language processing, computer vision and voice intelligence. In the scenario of Voice Intelligence Systems (VIS), PTMs have completely shifted the paradigm. From usage of handcrafted features and task-specific architectures towards rich, generalizable representations that encode acoustic, linguistic, and paralinguistic cues in a unified feature space. As a result, enabling downstream tasks to be learned efficiently even in low-resource or label-scarce settings. In this thesis, we focus on two critical frontiers of VIS: firstly, the affective dimension, which concerns the system's ability to interpret users' emotional and expressive states. Secondly, the secure dimension, which requires robustness against synthetic-speech manipulations such as voice cloning and the capability to attribute a fake to its generating model. Both of these tasks are essential for preserving trust and system integrity. These two capabilities are complementary and increasingly necessary across modern applications. Affective understanding supports conversational agents, healthcare monitoring, and human–computer interaction, whereas security underpins voice authentication, forensic analysis, and safety-critical communication. In emerging personalized and identity-aware services, both affective insight and resilience to synthetic speech are essential for trustworthy, user-centric VIS. Both dimensions have seen substantial progress with the rise of large PTMs, and recent research has increasingly leveraged these models to advance VIS. However, PTMs are highly heterogeneous—differing in architectural design, training objectives, linguistic scope, and modality grounding. Consequently, each PTM encodes a distinct and partial view of the speech signal, emphasizing different acoustic, paralinguistic, or semantic cues.

Recognizing this, researchers in application such as automatic speech recognition

have begun exploring fusion of heterogeneous PTMs representations to exploit complementary strengths for better performance, with similar trends emerging in different NLP and computer vision applications. In this thesis, we explore and deepen this direction by investigating how heterogeneous PTM representations can be systematically fused to advance both the affective and secure frontiers of VIS. To enable effective fusion, we introduce Proximal Interpolative Fusion, a novel fusion paradigm. Proximal refers to bringing diverse PTM representations into task-relevant compatibility across geometric, divergence, and interaction spaces. Interpolative denotes learning an intermediate fused manifold rather than collapsing one PTM's representational space into another, thereby preserving complementary paralinguistic, acoustic, and semantic cues. Building on this paradigm, we develop a set of novel fusion frameworks that are organized into three core families, each capturing a distinct principle for how heterogeneous representations are combined. (i) Geometry Guided Alignment enables fusion by placing PTM representations into shared geometric spaces prior to integration. This family includes **HYFuse**, which fuses representations through hyperbolic projection and möbius operations; **MATA**, which employs optimal transport for manifold-to-manifold alignment; and **MERLINN**, which performs dual-geometry fusion across euclidean and hyperbolic spaces. (ii) Interaction Guided Alignment achieves fusion through explicit interaction operators applied between representations. This includes **MiO**, which integrates PTM representations through bilinear pooling, followed by **SCAR**, which employs a cascaded cross-attention mechanism. (iii) Divergence Guided Alignment fuses representations by harmonizing their underlying distributional structure. This includes **FINDER**, which aligns representational distributions using renyi divergence, and **COFFE**, which leverages chernoff distance to jointly optimize separability. Additionally, **PARROT** serves as a hybrid framework that combines Interaction Guided with Geometry Guided Alignment through a hadamard interaction branch followed by optimal transport. Across affective and secure VIS tasks—including emotion recognition (**MATA**, **HYFuse**, **PARROT**) and synthetic-speech detection and attribution (**MiO**, **MERLINN**, **FINDER**, **SCAR**, **COFFE**), these frameworks consistently demonstrate that carefully structured fusion outperforms individual PTMs, underscoring its importance for developing improved affective and secure VIS in the era of PTMs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

VIS	**V**oice **I**ntelligence **S**ystems
SER	**S**peech **E**motion **R**ecognition
SSD	**S**ynthetic **S**peech **D**etection
CNN	**C**onvolutional **N**eural **N**etwork
PTM	**P**re-**T**rained **M**odel
SOTA	**S**tate-**o**f-**t**he-**a**rt
SSL	**S**elf-**S**upervised **L**earning
NVER	**N**on-**V**erbal **V**ocalization **E**motion **R**ecognition
HMM	**H**idden **M**arkov **M**odels
FCN	**F**ully **C**onnected **N**etwork
MFCC	**M**el-**F**requency **C**epstral **C**oefficients
GMM	**G**aussian **M**ixture **M**odel
RNN	**R**ecurrent **N**eural **N**etwork
SVM	**S**upport **V**ector **M**achine
MMS	**M**assively **M**ultilingual **S**peech
TDNN	**T**ime **D**elay **N**eural **N**etwork
NAC	**N**eural **A**udio **C**odec
RVQ	**R**esidual **V**ector **Q**uantization
DAC	**D**escript **A**udio **C**odec
IB	**I**mage**B**ind
LB	**L**anguage**B**ind
OT	**O**ptimal **T**ransport
EER	**E**qual **E**rror **R**ate
CFs	**C**odec**F**ake
CKA	**C**entered **K**ernel **A**lignment
TTS	**T**ext-**T**o-**S**peech
VC	**V**oice **C**onversion
SSSA	**S**ynthetic **S**peech **S**ource **A**ttribution

OT	**O**ptimal **T**ransport
LALMs	**L**arge **A**udio **L**anguage **M**odels
CFK	**C**odec**F**ake-**K**ids
CCFD	**C**hildren **C**odec**F**ake **D**etection
CKA	**C**entered **K**ernel **A**lignment
NNCES	**N**on-**N**ative **C**hildren's **E**nglish **S**peech
PCA	**P**rincipal **C**omponent **A**nalysis
RD	**R**ényi **D**ivergence
CD	**C**hernoff **D**istance
SSM	**S**tate-**S**pace **M**odel

CHAPTER 1

INTRODUCTION

1.1 Overview

"The voice is mirror to your soul"

The human voice—often described as a mirror to the soul—conveys far more than the literal meaning of spoken words. Embedded within its pitch, rhythm, timbre, and dynamics are rich paralinguistic cues that reflect a speaker's emotional state, intentions, identity, and communicative style. Over the past decade, voice has emerged as one of the most expressive and information-rich modalities for human–computer interaction, offering deep insight into the human condition. Although it is easy for humans to recognize and understand these cues, making it possible for computational system to do so with similar refinement is still a major problem-one that is at the core of VIS. These days, VIS are used for more than just assistive applications; they are the backbone of secure authentication, individualized services, remote healthcare, and seamless communication. In this thesis, we address two key dimensions of VIS: the affective dimension, which enables the system to interpret users' emotional and expressive states, and the secure dimension, which demands robustness to synthetic-speech attacks such as voice cloning and the ability to attribute a fake to its underlying generative model—both crucial for maintaining trust and system integrity. These capabilities complement one another and are increasingly vital across modern applications. Affective understanding drives conversational agents, healthcare monitoring, and human–computer interaction, while security forms the foundation of voice authentication, forensic investigation, and safety-critical communication. As personalized and identity-aware services continue to expand, the combination of emotional awareness and resilience to synthetic speech becomes essential for building trustworthy, user-centric VIS (see Figure 1.1). Within the affective dimension, we focus on tasks such as SER and NVER, both of which aim to infer users' emotional states—across the general population as well as for individuals with minimal speaking capabilities. Within the secure dimension, we study SSD,

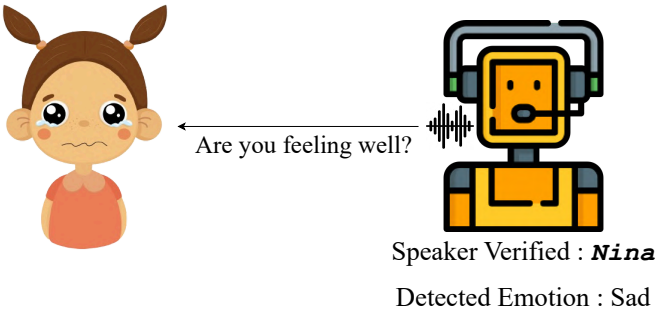


Figure 1.1: Affective and Secure VIS

including the edge case where the emotional content of speech is intentionally manipulated, followed by SSSA, which identifies the generative model responsible for producing a fake. Together, these tasks reflect the core requirement of affective and secure VIS: systems that not only process speech signals, but also understand and safeguard the users against synthetic manipulation.

1.2 Background

Over the years, substantial research within affective and secure VIS—particularly for tasks such as SER and SSD—has explored handcrafted features with classical learning methods Kishore and Satish (2013); Jin *et al.* (2015); Villalba *et al.* (2015); Balamurali *et al.* (2019). These approaches were later succeeded by deep learning architectures, including CNNs, capsule networks, and end-to-end neural models Issa *et al.* (2020); Qian *et al.* (2016); Luo *et al.* (2021); Ma *et al.* (2021). More recently by the dawn of this decade, the scenario of these tasks has been fundamentally changed by the rapid emergence of large-scale PTMs. Models such as Wav2vec2 Baevski *et al.* (2020), HuBERT Hsu *et al.* (2021*b*), and WavLM Chen *et al.* (2022*a*) learn rich acoustic, semantic, and paralinguistic representations from massive and diverse corpora, enabling strong performance even in low-resource settings. By eliminating the need to train task-specific models from scratch, PTMs significantly reduce both development time and computational cost. These models have been successfully adopted for SER Pepino *et al.* (2021); wen Yang *et al.* (2021) and SSD Eom *et al.* (2022), with further studies showing that augmenting PTM-based features with prosodic or pronunciation cues can improve robustness Wang *et al.* (2023). Similar trends are observed in NVER and SSSA, where earlier work relied on classical algorithms such as SVMs, followed by deep learning

architectures including residual networks and LSTMs Hsu *et al.* (2021*a*); Klein *et al.* (2024). More recent studies show that the incorporation of PTMs has substantially advanced these tasks as well Tzirakis *et al.* (2023); Klein *et al.* (2024); Xuan *et al.* (2025). The effectiveness of PTMs across affective and secure VIS stems from the highly informative and discriminative representations learned during large-scale pre-training, enabling downstream[1] models to achieve strong performance.

1.3 Research Gap and Motivation

The PTMs are highly heterogeneous, as they differ in architectural design (transformer- or state-space–based), pre-training objectives (self-supervised or supervised), linguistic scope (monolingual vs. multilingual), and modality grounding (multimodal vs acoustic-only) Mohamed *et al.* (2022). These variations create different inductive biases in the learned representational space, which means that each PTM encodes a different view of the speech signal, often with a non-overlapping description. They may encode different mixes of acoustic detail, semantic cues, and paralinguistic information Pasad (2025). This uniqueness among PTMs is not only a cause of inconsistency; it can also be an advantage. The variety in their learned representations could offer them strength in diversity through fusion-based methods. Recognizing this complementarity, research in ASR Arunkumar *et al.* (2022); Huang *et al.* (2023); Zhai and Zheng (2024) have explored fusion[2] of heterogeneous PTMs representations for improved performance. Moreover, such fusion of PTMs representations has also demonstrated effectiveness across a wide range of NLP, computer vision, multimodal learning applications Sharma *et al.* (2023); Don-Yehiya *et al.* (2023); Wan *et al.* (2024); Kesarwani *et al.* (2024); Chung *et al.* (2025). However, a principled fusion paradigm tailored to affective and secure VIS remains underexplored, leaving a critical gap in how heterogeneous PTM representations can be systematically aligned and integrated to improve performance. In this thesis, we work towards solving this research gap. However, existing research in speech processing for PTMs representations fusion mostly follow simple concatentation as fusion appraoch Arunkumar *et al.* (2022). While such approaches have demonstrated effectiveness, they often overlook the substantial representational heterogeneity

[1] *Throughout this thesis, we refer to downstream as any task-specific network or classifier attached after the PTM representations

[2] *Throughout this thesis, we use 'alignment' and 'fusion' interchangeably

that exists across PTMs. This raises the central question addressed in this thesis: *Can heterogeneous PTM representations be fused more effectively by explicitly accounting for the incompatibilities that exist between their representation spaces?* In this thesis, we work towards answering this research question.

1.4 Contributions

We introduce a novel paradigm, Proximal Interpolative Fusion, for the effective fusion of heterogeneous PTM representations. Unlike simple concatenation or linear fusion, this viewpoint emphasizes bridging representational gaps before integrating them, which becomes especially crucial when PTMs differ in architecture, pre-training objective, and modality grounding. In this paradigm, *proximal* refers to bringing diverse PTM representations into task-relevant compatibility across geometric, divergence, and interaction spaces, while *interpolative* denotes learning an intermediate fused manifold rather than collapsing one PTM's representational space into another. This design preserves complementary paralinguistic, acoustic, and semantic cues that each PTM uniquely contributes. Building on this paradigm, we develop a set of novel fusion frameworks that are organized into three core families, each capturing a distinct principle for how heterogeneous representations are combined for improved affective and secure VIS. Firstly, (i) Geometry Guided Alignment which enables fusion by embedding heterogeneous PTM representations into shared geometric spaces before integration, making it particularly suitable for tasks where manifold structure or hierarchical relationships are central. Within this family, **`HYFuse`** projects PTM representations into hyperbolic space and fuses them using mobius operations, preserving hierarchical structure across heterogeneous PTM representations for improved SER. **`HYFuse`** assumes the presence of inherent hierarchy among the information learned across the representations and exploits this structure for effective fusion. Following this, **`MATA`** employs optimal transport for manifold-to-manifold alignment across multimodal PTMs such as LB and IB for improved NVER. **`MATA`** assumes that heterogeneous PTMs contain complementary information residing on different manifolds and that preserving the distributional geometry of one space while aligning it to another enables more effective fusion. Lastly, **`MERLINN`** extends the geometric perspective introduced in **`HYFuse`** by performing dual-geometry fusion across euclidean and hyperbolic spaces, targeting

SSD and introducing the first child-speech–specific SSD dataset to expose vulnerabilities in current SSD systems. **MERLINN** assumes the existence of complementary properties of flat and curved geometric representational spaces and jointly exploits both for improved fusion. Secondly, (ii) Interaction Guided Alignment integrates representations by modeling explicit cross-representation interactions, capturing how different PTMs relate to and complement one another. Within this family, **MiO** applies bilinear pooling to integrate diverse PTM representations for SSD, while **SCAR** employs nested cross-attention to detect emotion-manipulated synthetic speech. **MiO** assumes that fine-grained interaction among the representations are needed for downstream performance while being efficient and simple. Similarly, **SCAR** assumes that explicitly modeling the fine-grained interactions in a nested manner leads to better performance. Thirdly, (iii) Divergence Guided Alignment fuses PTM representations by aligning their underlying distributional behavior. This family assumes that heterogeneous PTMs may encode related information using different distributional characteristics and that reducing such distributional discrepancies prior to fusion leads to better modeling of complementary information. **FINDER** employs renyi divergence to align PTM distributions for SSSA. In contrast, **COFFE** leverages chernoff distance for SSSA, focusing specifically on scenarios when the system is exposed to singing voices. Finally, **PARROT**, a hybrid framework that combines both Interaction Guided Alignment and Geometry Guided Alignment principles. **PARROT** integrates a hadamard interaction branch with optimal transport alignment to synergize heterogeneous PTM representations for SER. **PARROT** assumes that multiple forms of representational heterogeneity may coexist simultaneously. Collectively, these frameworks demonstrate how heterogeneous PTM representations can be effectively fused, establishing a unified roadmap for building next-generation affective and secure VIS.

1.5 Why not model the affective and secure dimension of VIS together?

We have introduced several novel frameworks following the paradigm of Proximal Interpolative Fusion to advance both the affective and secure dimensions of VIS. However, a natural question arises: since both dimensions rely heavily on paralinguistic

cues, why not model them jointly within a multi-task framework to streamline workflow, reduce computational cost, and simplify system maintenance? Although appealing in principle, such integration is neither practical nor beneficial in this context. Joint modeling risks negative transfer, where features useful for one task impair performance on the other. Furthermore, no large-scale dataset exists with both affect and secure dimension annotations, and curating one would be costly and complicated by licensing and privacy constraints. For these reasons, this thesis adopts separate, task-specific pipelines for affective and secure VIS, enabling each to be optimized independently while preserving the flexibility to integrate them in a two-stage or on-demand framework for specific downstream scenarios. For instance, in a customer support setting, only affective dimension may be required to monitor a caller's emotional state in real time, as the speech source is already trusted. In contrast, a mental health teleconsultation might require secure dimension followed by affective dimension—first verifying the authenticity of the voice to prevent impersonation, then assessing the patient's emotions for clinical insight. Conversely, scenarios like voice-based banking authentication demand only the secure dimension, where detecting synthetic speech is the sole priority and affect recognition is irrelevant. Accordingly, the thesis structures its contributions around these two complementary dimensions, each addressing distinct methodological challenges while collectively enabling the development of truly affective and secure VIS.

1.6 Organization

The thesis is organized as follows::

- Chapter 2: Related Works surveys prior research across affective and secure VIS. It traces the historical development of the field. It also provides an overview of the SOTA PTMs and highlights their architectural, objective, and modality-specific heterogeneity. Lastly, motivates how and why current PTM-based approaches can be strengthened through systematic fusion for affective and secure VIS.
- Chapter 3: Datasets describes the benchmark datasets used in this thesis.
- Chapter 4: Geometry Guided Alignment presents **`HYFuse`**, **`MATA`**, and **`MERLINN`**, three frameworks that align PTMs through geometric principles such as hyperbolic projection and optimal transport. This chapter demonstrates how manifold alignment enables complementary PTM representations to be fused more effectively.

- Chapter 5: Interaction Guided alignment introduces **MiO** and **SCAR**, two frameworks that employ explicit interaction operators—bilinear pooling and nested cross-attention—to model relationships between heterogeneous PTM representations.
- Chapter 6: Divergence Guided alignment introduces **FINDER** and **COFFE**, demonstrating how to align PTM representations by aligning the inner distributions using renyi divergence and chernoff distance, respectively.
- Chapter 7: Combination of Interaction Guided with Geometry Guided alignment introduces **PARROT**, a hybrid fusion framework combining mamba-based PTMs and transformer-based PTMs using a hadamard interaction branch followed by optimal transport.
- Chapter 8: Conclusions, Future Work and Publications summarizes the key insights derived from Proximal Interpolative Fusion and outlines several promising directions for future research in affective and secure VIS. Also, the publications carried out as part of this thesis are outlined.

CHAPTER 2

Related Works

This chapter reviewed prior work across the affective and secure dimensions of VIS, tracing the transition from traditional machine-learning approaches to neural network-based approaches and, more recently, to usage of PTMs on large and diverse speech corpora. This progression highlights the growing reliance on heterogeneous PTMs and motivates the need for a unified and principled fusion framework, which the subsequent chapters of this thesis aim to develop. We also provide overview of the SOTA PTMs which is later used in the succeeding chapters. All selected PTMs are publicly available and represent SOTA performance within their respective domains.

2.1 Affective Dimension

Early studies on SER relied on HMMs Cowie *et al.* (2001); Nogueiras *et al.* (2001); Schuller *et al.* (2003) and later on classical machine learning with handcrafted features Lee and Narayanan (2005). Following the success of AlexNet in ImageNet, CNN-based approaches gained popularity Huang *et al.* (2014), including hybrid designs with unsupervised and semi-supervised learning. Variants such as Gaussian Mixture-based HMMs, Subspace Gaussian Mixture HMMs, and deep learning-based HMMs were explored Mao *et al.* (2019). CNN architectures (e.g., AlexNet, ResNet, Xception, VGG, DenseNet) pre-trained on ImageNet have been applied to SER Ottl *et al.* (2020), with extensions integrating BiLSTM and attention Zhang *et al.* (2021) or fusing MFCCs with CNN features as inputs to LSTMs Araño *et al.* (2021*a*). While CNNs dominated early deep learning approaches, transformer-based models have emerged Lian *et al.* (2021); Lei *et al.* (2022); Wang *et al.* (2021), with Vision Transformers also applied to emotion recognition Heracleous *et al.* (2022). More recently, SER has shifted toward leveraging representations from PTMs such as YAMNet and Wav2vec Keesing *et al.* (2021), offering efficiency and improved performance by avoiding full model training and benefiting from large-scale, diverse pretraining. Self-supervised speech PTMs (WaLM, Unispeech-SAT, HuBERT) have outperformed other audio PTMs (e.g., YAMNet, VGGish) for SER, as shown in comparative evaluations Atmaja and Sasou (2022) using

FCN on top of extracted representations. Further, models trained on a single corpus often experience a notable decline in performance when deployed on previously unseen datasets, primarily due to variations in language, speaker characteristics, recording environments, and annotation practices. To improve robustness under such distribution shifts, recent research has increasingly focused on developing models capable of zero-shot (ZS) generalization across corpora Parry *et al.* (2019); Ahn *et al.* (2021); Agarla *et al.* (2024). In this context, language–audio pretrained models, particularly those built upon the Contrastive Language–Audio Pretraining (CLAP) framework Elizalde *et al.* (2023, 2024), have shown promising transfer capabilities by learning a shared embedding space that aligns audio and textual descriptions through contrastive objectives. A key advantage of CLAP-based models lies in their reliance on natural language supervision, enabling learning in a continuous semantic space rather than being constrained by predefined emotion labels. Nevertheless, the original CLAP architecture was developed primarily for audio–text retrieval and does not explicitly model emotional information. Consequently, affective characteristics are often entangled with other acoustic factors, limiting its effectiveness for emotion-related tasks. To overcome this challenge, several adaptations of CLAP have recently been proposed for SER Pan *et al.* (2024); Lin *et al.* (2024); Jing *et al.* (2024); Shi *et al.* (2025), introducing emotion-aware alignment objectives and affective representations to better capture emotional semantics. Among these approaches, ParaCLAP Jing *et al.* (2024) has gained particular attention, achieving state-of-the-art zero-shot performance across multiple SER benchmarks as well as other paralinguistic speech processing tasks.

2.2 Secure Dimension

ASVspoof 2015 corpus Wu *et al.* (2015*a*) marked a major milestone for SSD, catalyzing sustained research interest in the field. Early approaches primarily relied on GMMs and SVMs with handcrafted statistical features Sahidullah *et al.* (2015). Subsequent efforts transitioned toward neural architectures, including CNNs and RNNs Tom *et al.* (2018); Gomez-Alanis *et al.* (2019); Alzantot *et al.* (2019), followed by the adoption of SSL techniques Lee *et al.* (2023); jin Shim *et al.* (2020); Jiang *et al.* (2020). Further, a wide range of speech PTMs—such as Mockingjay, TERA, Wav2vec, HuBERT—have been evaluated for SSD Eom *et al.* (2022); wen Yang *et al.* (2021). Notably, Wang

et al. (2023) demonstrated that integrating Wav2vec representations with prosodic and pronunciation cues significantly improves generalization. More recently, the works of Wu *et al.* (2024*b*) and Lu *et al.* (2024) laid the groundwork for NAC–generated speech detection, a direction motivated by the rapid rise of LALMs that employ NACs in their synthesis pipelines. These studies revealed that SOTA SSD models—largely trained on vocoder-based deepfake datasets—experience substantial performance degradation when evaluated on NAC-generated speech. Building on this, Lu *et al.* (2024) expanded dataset construction by incorporating both VCTK and AISHELL3 corpora and evaluated AASIST and LCNN architectures using mel-spectrogram and Wav2vec2-based features. Subsequently, Chen *et al.* (2025*a*) further extended this line of work by introducing a wider variety of NACs for synthetic speech generation, again leveraging AASIST with Wav2vec2 representations to improve detection robustness. More recently, Chen *et al.* (2025*b*) proposed SASTNet, a unified framework that combines semantic representations from Whisper with acoustic features from Wav2vec2 and AudioMAE for NAC-generated speech detection. While these approaches primarily target NAC–generated speech detection, they do not explicitly address vocoder-based synthesis. To bridge this gap, Xie *et al.* (2025*a*) introduced a unified detection strategy capable of handling both CodecFake and vocoder-synthesized speech using sharpness-aware minimization.

As the field of SSD matures, the research focus is gradually expanding beyond conventional binary detection toward SSSA, where the objective is not only to determine whether an audio sample is synthetic but also to identify the specific generative system responsible for its creation, such as a NAC, TTS, or VC model. This shift marks an important evolution in audio forensic research, as source-level analysis offers greater interpretability, transparency, and practical value for real-world deployment and regulatory applications. To address this emerging challenge, a variety of SSSA methods have been proposed in recent years Yan *et al.* (2022*a*); Chhibber *et al.* (2025). More recently, attention has turned toward open-set source attribution, where the system must operate in the presence of previously unseen generators encountered during inference Bhagtani *et al.* (2024); Xie *et al.* (2024); Wang *et al.* (2025). Advancing this line of work, Xie *et al.* 2025*b* introduced a source attribution framework for CodecFakes that is capable of handling both in-distribution and open-set conditions, addressing a critical challenge in the rapidly evolving area of synthetic speech forensics.

2.3 Pre-Trained Models

The first subsection discusses PTMs trained for speech representation learning, covering monolingual, multilingual, and speaker-recognition-models. The second subsection introduces NACs, which have recently gained focus in the community as foundational backbones for PTM-based speech processing. The third subsection presents multimodal PTMs developed through multimodal pretraining. Lastly, we describe music PTMs trained for musical representation learning, which are employed in SSSA to assess performance on synthetic singing voice (Chapter 6). Finally, we provide a separate discussion on Audio-Mamba, a SSM-based PTM. It is utilized in Chapter 7 to examine fusion strategies that arise specifically from architectural heterogeneity. We present it independently from the preceding PTMs because, unlike the transformer-based architectures dominating prior sections, Audio-Mamba constitutes state-space modeling.

2.3.1 Speech PTMs

Monolingual: Monolingual PTMs taken under consideration are—Unispeech-SAT, WavLM, Wav2vec2, and HuBERT—all primarily trained on the English LibriSpeech corpus, with WavLM (Large) additionally consisting of Mix 94k dataset. Unispeech-SAT Chen *et al.* (2022*b*) integrates contrastive learning and multitask objectives within a speaker-aware pre-training framework; we use its base variant, which contains 94.68M parameters. WavLM Chen *et al.* (2022*a*) jointly optimizes masked speech prediction and denoising tasks, enabling it to capture both phonetic content and speaker traits. Our study includes the base (94.70M parameters) and large (316.62M parameters) configurations. Wav2vec2 Baevski *et al.* (2020) performs mask-based representation learning using a contrastive loss over quantized speech targets; we adopt the base model comprising 95.04M parameters and 12 transformer encoder layers. HuBERT Hsu *et al.* (2021*b*), with 94.68M parameters, follows a BERT-style masked prediction objective.

Multilingual: The multilingual PTM suite includes XLS-R Babu *et al.* (2022), Whisper Radford *et al.* (2023), and MMS Pratap *et al.* (2023), which collectively span 128, 96, and more than 1400 languages. XLS-R is built upon Wav2vec2 as backbone and is trained in a self-supervised manner on 436k hours of multilingual audio sourced from VoxPopuli, CommonVoice, Multilingual LibriSpeech, VoxLingua107, BABEL, and re-

lated corpora. Its learning framework follows a contrastive prediction objective over masked encoder representations. In our experiments, we utilize both the 1B-parameter and 300M-parameter variants. Whisper adopts an encoder–decoder formulation and is trained using weak supervision across 680k hours of web-collected speech–text data, optimized for multilingual transcription and related tasks. We employ the base configuration containing 74M parameters. MMS likewise builds on the Wav2vec2 family but combines a convolutional front-end with transformer layers. Pre-training was done in a constrastive manner on more than 500k hours of speech from MMS-lab, FLEURS, BABEL, and additional multilingual datasets, offering the broadest language coverage among the models considered. We use the publicly released 1B-parameter model.

Speaker Recognition: x-vector Snyder *et al.* (2018*a*) is speaker recognition PTM trained on VoxCeleb1 and VoxCeleb2 datasets. x-vector is a SOTA speaker recognition system based on a TDNN trained end-to-end in a supervised fashion without relying on handcrafted features. Once trained, it generates fixed-dimensional speaker representations from variable-length utterances and outperforms the earlier i-vector approach. For our work, we use the publicly available implementation of x-vector from *HuggingFace* via the *SpeechBrain* Ravanelli *et al.* (2021) library.

Usage of the Speech PTMs in this Thesis: For the speech PTMs, the inputs are resampled to 16 kHz before being fed into the respective PTMs. Feature extraction is performed by applying average pooling over the final hidden states of each frozen PTM; for Whisper, representations are obtained from the encoder. The resulting embedding dimensions are 768 for WavLM (Base), Wav2vec2, and Unispeech-SAT; 512 for Whisper and x-vector; 1024 for WavLM (Large); and 1280 for both XLS-R and MMS.

2.3.2 Music PTMs

The MERT family of models Li *et al.* (2023) represents SOTA line of music PTMs, designed to capture rich and fine-grained structure from musical audio. Owing to their large-scale pretraining on diverse music corpora, these models achieve strong performance across numerous music processing tasks. In this thesis, we employ several variants of MERT—MERT-v1-330M, MERT-v1-95M, MERT-v0-public, and MERT-v0. We additionally include music2vec-v1Li *et al.* (2022), a self-supervised music PTM

that learns generalized representations suitable for a wide range of music information retrieval tasks. Except for MERT-v1-330M, which comprises 330M parameters, all the other music PTMs considered have approx 95M parameters.

Usage of music PTMs in this Thesis: MERT-v0-public, MERT-v0, and music2vec-v1 process input at 16 kHz whereas MERT-v1-330M and MERT-v1-95M operate on 24 kHz. The representations are extracted by applying average pooling to the final hidden layer of each frozen PTM. All msuic PTMs yield 768-dimensional representations, with the exception of MERT-v1-330M (1024-dimension).

2.3.3 Neural Audio Codecs

We provide a more detailed treatment of these models because they play a dual role in our work: (i) their encoder representations are used for SER within Chapter 4 (**`HYFuse`**), and (ii) they serve as the generative backbone for synthesizing children's voices in Chapter 4 (**`MERLINN`**).

DAC Kumar *et al.* (2024) is a high-quality NAC based on the VQ-GAN framework. It incorporates residual vector quantization along with adversarial and multi-scale spectral reconstruction losses to achieve superior audio fidelity. For our experiments, we utilize multiple DAC configurations corresponding to three sampling rates: 16 kHz, 24 kHz, and 44 kHz.

EnCodec Défossez *et al.* (2022) is a real-time NAC designed for high-fidelity compression. It uses a convolutional encoder–decoder architecture combined with RVQ and is trained using a mix of time-domain, frequency-domain, and spectrogram-based adversarial losses to improve reconstruction quality. In this thesis, we adopt both 24 kHz and 48 kHz versions of EnCodec.

SoundStream Zeghidour *et al.* (2021) is a NAC designed for efficient low-bitrate speech compression. It employs an encoder–decoder architecture with RVQ and multi-scale STFT discriminators, enabling it to maintain perceptual quality while operating at bitrates ranging from 3 to 18 kbps. For our experiments, we use the 16 kHz model variant.

SpeechTokenizer Zhang *et al.* (2024*b*) is a unified NAC framework that aims to connect acoustic and semantic representations for speech language modeling. It follows an

encoder–decoder design with RVQ, where different layers progressively disentangle linguistic content from paralinguistic characteristics. In our experiments, we adopt the standard configuration, which produces a combined token sequence across all layers, and use the 16 kHz model variant.

FunCodec Du *et al.* (2024) combines RVQ with semantic enhancement and adversarial learning mechanisms to produce compact yet expressive discrete audio units. For our experiments, we use the officially released FunCodec models trained on LibriTTS and large-scale bilingual speech corpora at a 16 kHz sampling rate.

AudioDec Wu *et al.* (2023*c*) is a high-fidelity NAC based on a two-stage autoencoder pipeline. Initially, the model is optimized using metric-based losses to ensure stable encoder–decoder convergence. A subsequent adversarial training phase—applied only to the decoder—further refines waveform realism. In this work, we utilize AudioDec variants operating at 28 kHz and 48 kHz.

SNAC Siuzdak *et al.* (2024) augments the standard residual vector quantization paradigm by incorporating a hierarchy of quantizers functioning at multiple temporal scales. Its architecture integrates depthwise convolutions for robust training, noise-injection modules to enhance representational expressiveness, and localized windowed attention to preserve contextual structure. We use several SNAC variants in our experiments, including models trained at 24 kHz, 32 kHz, and 44 kHz sampling rates.

2.3.4 Multimodal PTMs

IB Girdhar *et al.* (2023) is a multimodal PTM that jointly learns from images, text, audio, depth maps, IMU readings, and thermal data. Its design projects all non-visual modalities into a unified image-aligned embedding space. Built on a transformer backbone and optimized using an InfoNCE-style contrastive objective, IB supports zero-shot reasoning across modalities.

LB Zhu *et al.* (2023) adopts language as the central representational anchor, exploiting its rich semantic structure. The model aligns different modalities to a fixed language encoder through contrastive learning. Trained on the large-scale VIDAL-10M corpus, LB achieves SOTA performance across multiple multimodal benchmarks.

Usage of multimodal PTMs in this Thesis: All audio inputs are resampled to 16 kHz before being fed into the audio encoders of the multimodal PTMs. Feature representations are obtained by application of mean pooling over the final hidden layer of each PTM, yielding embeddings of 1024-dimension for IB and 768-dimension for LB.

2.3.5 Audio-Mamba

Audio-Mamba Yadav and Tan (2024) is a selective SSM PTM designed for self-supervised learning of universal audio representations. It operates by predicting randomly masked regions of the input spectrogram, enabling it to capture both local and long-range structure without relying on attention mechanisms. Pre-trained on the large-scale AudioSet corpus, Audio-Mamba has been shown to surpass several transformer-based baselines on a wide range of audio and speech tasks, including SER. In this thesis, we utilize the tiny (4.8M parameters), small (17.9M parameters), and base (69.3M parameters) variants.

Usage of Audio-Mamaba in this Thesis: Representations are extracted from the last hidden layer using average pooling, resulting in embedding dimensions of 960, 1920, and 3840 for the tiny, small, and base variants, respectively.

N.B: Over all the experiments in this thesis, we keep the PTMs frozen.

2.4 Fusion Paradigms, Research Gap & Takeaway

Multimodal learning has consistently been at the forefront of developing fusion paradigms, as its central objective is the effective integration of information from multiple modalities. Consequently, a wide range of fusion strategies has been proposed, ranging from simple feature concatenation to sophisticated attention-based and alignment-based fusion mechanisms Wang *et al.* (2020*b*); Gao *et al.* (2020); Steyaert *et al.* (2023); Khan *et al.* (2024); Zhao *et al.* (2024*a*). Simple concatenation implicitly assumes that the representations being fused are already compatible. To overcome this limitation, numerous studies have proposed more advanced fusion paradigms based on specific assumptions regarding the relationships between representations. For example, Kumar and Nandakumar (2022) employed outer-product fusion for multimodal hateful meme

classification, assuming that explicitly modeling fine-grained interactions between textual and visual representations can improve downstream performance. Similarly, Araño *et al.* (2021*b*) utilized hyperbolic geometry to model hierarchical relationships among visual, textual, and audio modalities for multimodal emotion recognition, assuming the existence of an underlying hierarchical structure across modalities.

However, these fusion paradigms were primarily developed for multimodal data, where each modality contributes fundamentally different information. In contrast, affective and secure VIS increasingly relies on heterogeneous PTMs operating on the same underlying speech signal. Consequently, the central challenge is not modality fusion, but representational heterogeneity arising from differences in architecture, pretraining objectives, linguistic scope, and modality grounding. Prior research in affective and secure VIS has extensively explored individual PTMs; however, the systematic fusion of PTM representations for improving downstream performance remains largely underexplored. Within speech processing, limited efforts in this direction have primarily appeared in speech recognition Arunkumar *et al.* (2022) and have largely relied on simple concatenation-based fusion.

Motivated by this gap, this thesis investigates PTM fusion through the proposed Proximal Interpolative Fusion paradigm, which explicitly accounts for heterogeneity across PTM representational spaces. This perspective naturally leads to Geometry Guided, Interaction Guided, and Divergence Guided Alignment paradigms. In parallel to this thesis, Wu *et al.* (2023*a*) explored simple concatenation-based fusion, which we adopt as a baseline across our SER studies (Chapter 4: **`HYFuse`** and Chapter 7: **`PARROT`**). In contrast, the proposed frameworks consistently outperform concatenation-based fusion, demonstrating the effectiveness of principled alignment-driven fusion strategies over naive representation combination methods.

CHAPTER 3

Datasets

In this chapter, we present the datasets used across the tasks considered within affective and secure VIS. The datasets are SOTA benchmark datasets across the tasks. In the first section, we present the dataset considered for tasks across the affective dimension and in the second section, along the secure dimension.

3.1 Affective Dimension

3.1.1 SER

Brief information regarding the datasets considered for SER used in Chapter 4 **HYFuse**, **MATA**, and Chapter 7 **PARROT**.

CREMA-D Cao *et al.* (2014): It is an English, gender-balanced dataset with 48 male and 43 female actors delivering 7,442 utterances. It contains a range of ages and ethnicities, expressing six emotions—anger, happiness, sadness, fear, disgust, and neutral—spoken from 12 scripted sentences.

BAVED Aouf (2019): It consists of 1,935 Arabic recordings from 45 male and 16 female speakers. It encompasses three emotional categories: neutral, low-intensity (e.g., tiredness), and high-intensity (e.g., happiness, sadness, anger).

URDU Latif *et al.* (2018): It contains 400 spontaneous utterances for anger, happiness, sadness, and neutral states, taken from unscripted television talk show conversations between 27 male and 11 female speakers.

Emo-DB Burkhardt *et al.* (2005): It is a German dataset that contains 535 utterances of five male and five female actors reading ten scripts. It consists of seven emotions: anger, anxiety, sadness, boredom, disgust, happiness and lastly neutral.

AESDD Vryzas *et al.* (2018):It is Greek dataset consisting of around 600 utterances from five actors, which holds five emotions—anger, disgust, fear, happiness, and sadness.

MESD Duville *et al.* (2021): It is a Mexican Spanish speech corpus comprising 864 utterances that covers: anger, sadness, fear, happiness, disgust and neutral.

3.1.2 NVER

Brief information regarding the datasets considered for NVER used in Chapter 4 **MATA**.

ASVP-ESD Landry *et al.* (2020): It contains thousands of high-quality recordings labeled with 12 emotions plus a breath class, captured in natural settings from diverse speakers. We use only the non-speech segments, sourced from films, TV, YouTube, and other media.

JNV Xin *et al.* (2023): It comprises 420 clips from four native Japanese speakers (two male, two female) expressing six emotions: anger, disgust, fear, happiness, sadness, and surprise. Recorded at 48 kHz in an anechoic chamber, it includes both scripted and spontaneous vocalizations.

VIVAE Holz *et al.* (2022): It consists of 1085 samples from eleven speakers covering three negative (anger, fear, pain) emotions and three positive (achievement, pleasure, surprise), recorded at varying intensities.

3.2 Secure Dimension

In this section, we provide brief information regarding the datasets considered in Chapter 4 **MERLINN**, Chapter 5 **MiO** & **SCAR**, and Chapter 6 **FINDER** & **COFFE**. These datasets are used in both SSD and SSSA.

ASVSpoof 2019 (ASV) Wang *et al.* (2020*a*): The dataset was created to support research on SSD and to strengthen automatic speaker verification systems against malicious manipulation. It includes three broad categories of spoofing attacks—TTS synthesis, VC, and replay—produced using SOTA neural acoustic and waveform generation techniques. The corpus contains both genuine (bonafide) recordings and synthetic speech, with the latter generated by 19 different synthesis systems. All audio files are provided at 16 kHz sampling rate. The bonafide speech features a diverse set of speakers reflecting variation in accent, speaking style, and vocal traits. For SSD (Chapter 5

MiO), we train, validate, and assess the model on the respective dataset partitions. For SSSA (Chapter 6 **FINDER**), we utilize all 19 generation systems as distinct classes. Dataset statistics are summarized in Table 3.1.

Category	Frequency
Dev	24844
Train	25380
Eval	71237
Total Samples	**121461**
Real	12483
Total Fake Samples	108978

Table 3.1: ASV Statistics

In-the-Wild Dataset (ITW) Müller *et al.* (2022*a*): This dataset contains approximately 37.9 hours of audio and includes recordings of well-known English-speaking public figures, such as celebrities and political leaders. It incorporates synthetic speech samples that mirror the types of deepfakes circulating in real-world settings, making the task of distinguishing authentic audio from manipulated content particularly challenging. The fake clips for each individual were collected from publicly available sources, including social media platforms and video-sharing websites. Owing to its realistic nature, the dataset is a valuable benchmark for evaluation of SSD systems under real-world conditions. Since no official train–validation–test partition is provided for ITW, we adopt a custom split: 70% for training, 10% for validation, and 20% for testing (Chapter 5 **MiO**).

DECRO Ba *et al.* (2023): SSD systems often struggle to generalize across languages, with models trained on one language performing poorly when evaluated in a zero-shot setting on another. To address this limitation, Ba *et al.* 2023 introduced the DECRO dataset, designed specifically to assess SSD performance in cross-lingual scenarios. DECRO contains both genuine and synthetic speeh samples in English (DECRO-E) and Chinese (DECRO-C). In this work, we follow the official train–validation–test partition provided by Ba *et al.* 2023 in for SSD (Chapter 5 **MiO**). The distribution of real and fake samples for each subset is illustrated in Figure 3.1.

FAD Chinese Dataset (CFAD) Ma *et al.* (2024): This dataset was developed to fill the gap in publicly available Chinese corpora for SSD. It contains both genuine speech and synthetic speech produced using 12 modern TTS synthesis and VC techniques.

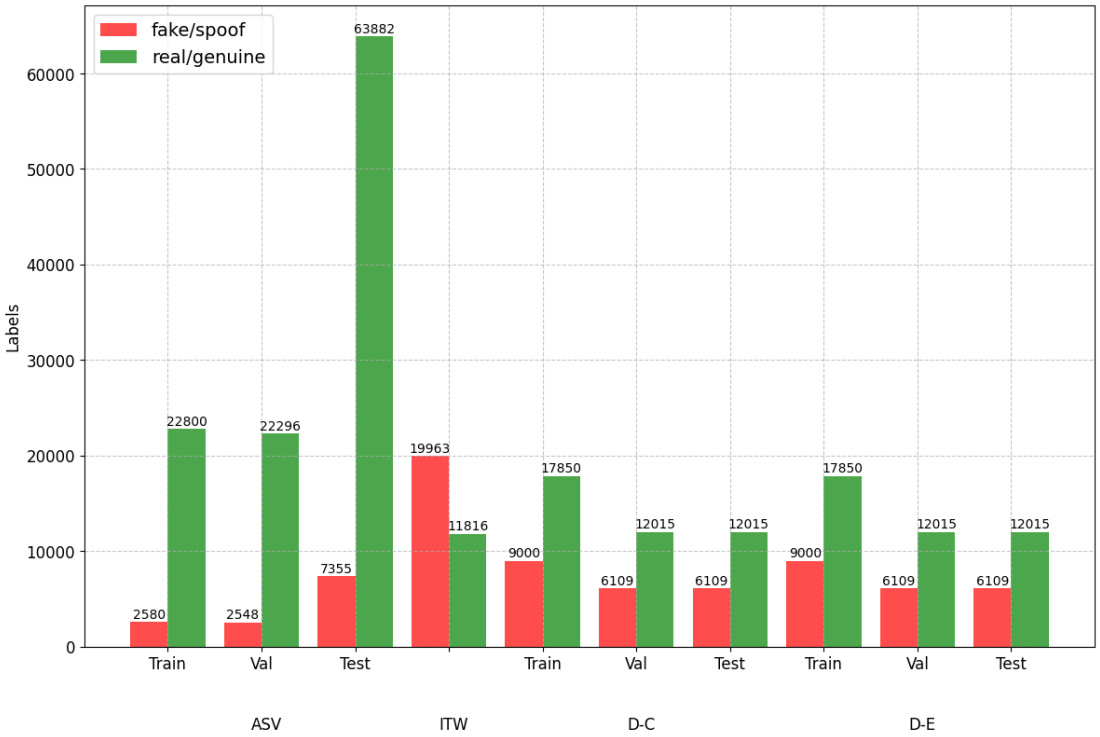


Figure 3.1: The distribution of samples by class across the datasets; D-C and D-E refer to the DECRO Chinese and English datasets respectively

To better reflect real-world scenarios, three different noise sources were incorporated at five signal-to-noise ratio levels. The synthetic recordings were generated using 11 SOTA TTS and VC systems.

Emofake Zhao *et al.* (2024*b*): We employ the only publicly available dataset for SSD that specifically targets cases where the emotional qualities of speech are intentionally modified. The corpus contains recordings in both English and Chinese, featuring multiple speakers expressing five core emotions: Neutral, Happy, Angry, Sad, and Surprise. The synthetic speech samples were generated using seven open-source emotional voice conversion systems, which alter affective cues while maintaining the speaker's identity. The dataset is divided into separate English and Chinese sections, each providing predefined training, development, and test sets. Both language partitions include 27300 training clips, 9100 development clips, and 17500 test clips.

CtrSVDD Zhang *et al.* (2024*c*): This dataset serves as a dedicated benchmark for detecting synthetic singing voices and contains audio samples in both Chinese and Japanese. We employ it for SSSA, focusing on scenarios in which the system must process singing voice inputs (Chapter 6: **COFFE**). The corpus consists of 220798 mono vocal clips (307.98 hours in total), categorized into bonafide singing and artificially

generated vocals. For our experiments, we exclusively use the synthetic portion, which includes 188486 clips amounting to 260.34 hours of audio. These spoofed samples were produced using 14 different vocal synthesis approaches (A01–A14). The dataset provides its own official train, development, and evaluation partitions.

CHAPTER 4

Geometry Guided Alignment

In this chapter, we introduce the Geometry Guided Alignment frameworks developed in this thesis. Prior work has demonstrated the utility of aligning representations via geometric-aware learning for multimodal fusion. For instance, Arano et al. Araño *et al.* (2021*b*) exploited the existence of hierarchical relationship between two different modalities by leveraging hyperbolic space to improve multimodal sentiment classification. In contrast, Pramanick et al. Pramanick *et al.* (2021) leveraged optimal transport to align heterogeneous modalities for sarcasm and humor detection. These efforts highlight that geometry guided fusion methods can provide smoother, more semantically meaningful alignment across modalities. However, despite these advances, geometry guidedfusion has not been explored for aligning heterogeneous PTM representations for improving affective and secure VIS or neither in the whole speech processing domain. This thesis addresses this gap by developing Geometry Guided Alignment methods and demonstrating their effectiveness across SER, NVER, and SSD tasks within affective and secure VIS.

4.1 HYFUSE

In this section, we present the background and motivation for employing hyperbolic space to fuse heterogeneous PTM representations, and subsequently introduce our proposed framework, **`HYFuse`** Phukan *et al.* (2025*d*), which harnesses this geometric formulation to achieve improved SER performance.

4.1.1 Background and Motivation

Recent SER advancements increasingly leverage representations from SOTA PTMs, which have driven notable performance gains. These representations fall into two broad categories: *representation-learning-based representations* (RLRs) from speech

PTMs such as Wav2vec2 Baevski *et al.* (2020), WavLM Chen *et al.* (2022*a*), and XLS-R Babu *et al.* (2022), and *compression-based representations* (CBRs) from neural audio codecs like EnCodec Défossez *et al.* (2022), DAC Kumar *et al.* (2024), and SoundStream Zeghidour *et al.* (2021). RLRs, trained for speech representation learning, capture higher-level prosodic features, while CBRs, learned via encoder–decoder compression, preserve fine-grained acoustic cues such as pitch and timbre. Previous works have explored RLRs Pepino *et al.* (2021); Morais *et al.* (2022) and, more recently, CBRs for SER Wu *et al.* (2024*a*); Ren *et al.* (2024); Mousavi *et al.* (2024), with studies showing gains from fusing multiple RLRs Wu *et al.* (2023*b*). Further, similar improvements have been observed across related task such as automatic speech recognition Arunkumar *et al.* (2022). However, no prior work has examined fusion of heterogeneous RLRs and CBRs. In this work, we present, to the best of our knowledge, the first investigation into fusing RLRs and CBRs for SER. *Our central hypothesis is that combining these two representations can further enhance SER performance by exploiting their complementary strengths: CBRs tend to capture fine-grained acoustic attributes such as pitch and timbre, whereas RLRs encode higher-level paralinguistic characteristics*. To our end, we introduce **`HYFuse`**, a novel framework for effective fusion of RLRs and CBRs that maps both representation types from euclidean space into hyperbolic space and performs fusion via mobius addition. To our knowledge, this is the first application of hyperbolic geometry for representation fusion in SER. **`HYFuse`** preserves the hierarchical relationship inherent in the two representation types and enables a more coherent alignment of low-level acoustic cues with higher-level paralinguistic patterns.

The key Contributions are as follows:

- We introduce **`HYFuse`** (Figure 4.1), a novel framework that maps RLRs and CBRs into hyperbolic space, exploiting its geometric properties for more effective fusion.
- With **`HYFuse`**, combining x-vector (RLRs) and SoundStream (CBRs) yields superior performance over individual and homogeneous fusion, achieving new SOTA results on CREMA-D and Emo-DB, demonstrating the benefits of fusion of heterogeneous PTMs representations for SER.

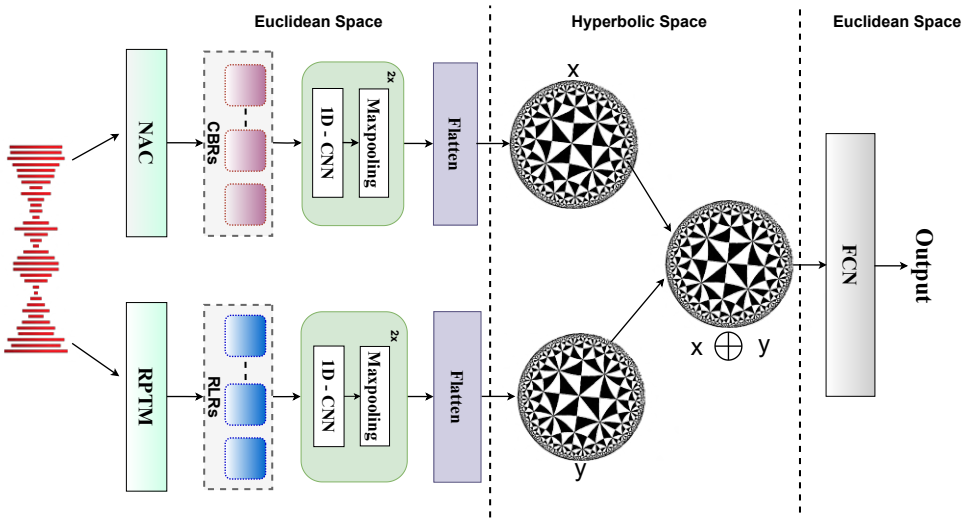


Figure 4.1: **HYFuse**; $x \oplus y$ stands for mobius addition

4.1.2 HYFuse: Framework

HYFuse (Figure 4.1) leverages hyperbolic geometry to integrate RLRs and CBRs while preserving their hierarchical relationships. This design captures both low-level acoustic cues (from CBRs) and high-level prosodic features (from RLRs), maintaining relative distances in the feature space, unlike euclidean space which may distort such structure. The architecture is shown in Figure 4.1. First, each representation is processed through the same 1D CNN block used in individual modeling (Below in Experiments: Modeling with Individual PTM representations). Flattened features $\mathbf{u}$ (from RLRs) and $\mathbf{v}$ (from CBRs) are then projected from euclidean to hyperbolic space using the exponential map:

$$\exp_{\mathbf{0}}(\mathbf{z}) = \begin{cases} \tanh(\lambda\|\mathbf{z}\|)\frac{\mathbf{z}}{\|\mathbf{z}\|}, & \|\mathbf{z}\| > 0, \\ \mathbf{0}, & \|\mathbf{z}\| = 0, \end{cases} \tag{4.1}$$

where $\lambda > 0$ is the curvature parameter and $\|\cdot\|$ is the euclidean norm. Let $\mathbf{u}_h$ and $\mathbf{v}_h$ be the hyperbolic projections. They are fused via mobius addition:

$$\mathbf{u}_h \oplus \mathbf{v}_h = \frac{(1 + 2\langle \mathbf{u}_h, \mathbf{v}_h \rangle + \|\mathbf{v}_h\|^2)\mathbf{u}_h + (1 - \|\mathbf{u}_h\|^2)\mathbf{v}_h}{1 + 2\langle \mathbf{u}_h, \mathbf{v}_h \rangle + \|\mathbf{u}_h\|^2\|\mathbf{v}_h\|^2}, \tag{4.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the euclidean dot product. The fused vector $\mathbf{w}_h$ is mapped back to euclidean space via the logarithmic map:

$$\log_{\mathbf{0}}(\mathbf{w}_h) = \begin{cases} 2 \cdot \operatorname{arctanh}(\|\mathbf{w}_h\|) \frac{\mathbf{w}_h}{\|\mathbf{w}_h\|}, & \|\mathbf{w}_h\| < 1, \\ \mathbf{0}, & \|\mathbf{w}_h\| = 0. \end{cases} \tag{4.3}$$

Finally, an FCN block followed by a softmax output layer predicts emotion classes. The trainable parameter count for **`HYFuse`** ranges from 8M to 13M, depending on the input representation pair.

Rep	CREMA-D				Emo-DB			
	FCN		CNN		FCN		CNN	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
RLRs								
W2V	58.66	55.14	65.16	65.08	88.21	86.42	**91.51**	**90.65**
WLM	64.71	61.63	68.81	68.64	87.23	85.21	89.72	89.52
XVC	63.69	61.20	68.77	68.67	85.65	85.25	81.31	80.61
HBT	67.85	66.25	70.63	70.45	86.25	85.39	88.81	87.85
CBRs								
EDC	47.85	46.98	48.48	42.56	47.65	45.08	48.04	40.20
DAC	42.52	41.86	43.84	35.74	39.78	38.52	40.56	34.10
STZ	41.61	40.14	48.51	45.92	45.65	43.71	49.91	37.20
SSM	54.20	53.94	55.19	53.03	59.12	57.96	**61.36**	**58.96**

Table 4.1: Performance comparison of models trained with different RLR and CBR feature sets. Results are reported as percentages and represent the average over five folds. "Acc" and "F1" denote accuracy and macro-averaged F1 score, respectively. The abbreviations used are: Wav2vec2 (W2V), WavLM (WLM), x-vector (XVC), HuBERT (HBT), EnCodec (EDC), DAC (DAC), SpeechTokenizer (STZ), and SoundStream (SSM). The same notation is maintained in Table 4.2 for consistency.

4.1.3 Experiments

Modeling with Individual PTM representations: For standalone RLRs or CBRs modeling, we employ a FCN and a CNN. The CNN consists of two 1D convolutional layers (64 and 128 filters, kernel size = 3) with ReLU activation, followed by flattening and a dense layer with 128 units. The FCN variant uses the same dense block without convolutional layers. The output layer in all cases applies a softmax activation for

emotion classification.

Dataset: We use CREMA-D and Emo-DB. Detailed above in Chapter 3.

Pre-Trained Representations: For RLRs, the process for extraction and dimension of the representations are detailed in Chapter 2 **Speech PTMs**. We use Wav2vec2, WavLM (Base), x-vector and HuBERT for our experiments. For the CBRs, all the audios is resampled to 16 kHz before being processed by DAC, Soundstream, and SpeechTokenizer; EnCodec operates at its native 24 kHz sampling rate. We extract codec-based representations from the frozen encoder outputs using average pooling. The resulting feature dimensionalities are 256 for Soundstream, 251 for DAC, 250 for SpeechTokenizer, and 375 for EnCodec.

Training Details: The models are optimized using Adam with a learning rate of 1e-5, batch size of 32, and trained for 50 epochs. Cross-entropy loss serves as the objective function. Dropout regularization is applied to prevent overfitting. Evaluation follows a 5-fold cross-validation scheme, where four folds form the training set and the remaining fold is used for testing.

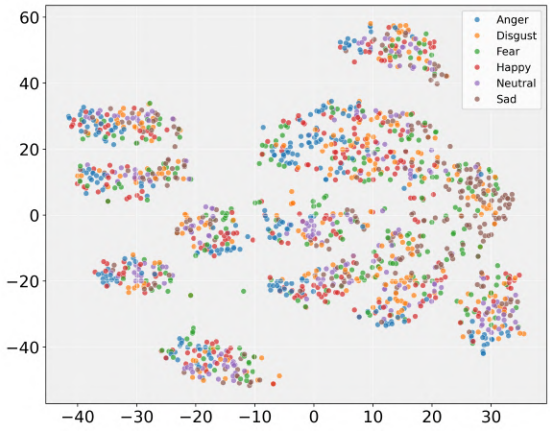
Experimental Results: Table 4.1 presents the performance of downstream models trained on individual RLRs and CBRs. As expected and consistent with prior work Mousavi *et al.* (2024), RLRs outperform CBRs across all SER datasets, reflecting the limited ability of CBRs alone to capture the paralinguistic cues crucial for SER. Among the CBRs, SoundStream yields the strongest results, though it still falls short of the best-performing RLRs. For the RLR-based models, HuBERT paired with a CNN achieves the highest accuracy on CREMA-D, while Wav2vec2 with CNN performs best on Emo-DB, illustrating that dataset-specific characteristics influence downstream performance. Across all settings, CNN classifiers consistently exceed FCN models. These results provide the baseline against which our fusion methods are evaluated.

Table 4.2 compares homogeneous fusions (RLR+RLR, CBR+CBR) and heterogeneous fusions (RLR+CBR). We use concatenation in euclidean space as the baseline fusion method, matching **`HYFuse`** in architecture and training settings for fairness. Results show that **`HYFuse`** consistently outperforms both individual models and concatenation across datasets, highlighting the advantage of hyperbolic fusion. Homogeneous RLR combinations, such as Wav2vec2+HuBERT, improve over individual PTM rep-

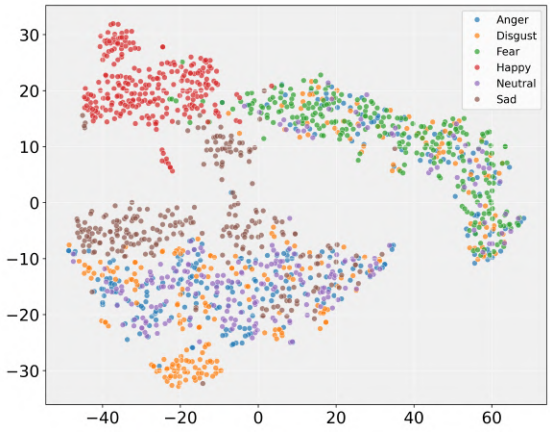
Pairs	CREMA-D				Emo-DB			
	Concat		HYFuse		Concat		HYFuse	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
RLRs + RLRs								
W2V + WLM	60.98	59.65	64.58	63.38	88.63	87.36	93.36	92.27
W2V + XVC	58.78	57.73	66.88	65.53	87.69	86.64	91.45	90.03
W2V + HBT	71.18	70.36	76.61	75.52	87.25	86.13	94.63	94.48
WLM + XVC	69.33	68.18	74.49	73.38	88.24	87.34	92.08	91.36
WLM + HBT	65.97	65.82	77.25	76.69	89.64	88.61	94.25	93.37
XVC + HBT	68.52	67.79	73.64	72.28	88.33	87.79	93.62	93.57
CBRs + CBRs								
EDC + DAC	58.97	47.61	64.68	63.26	58.96	52.45	64.85	63.28
EDC + STZ	58.97	45.08	66.67	65.59	58.68	56.51	64.14	63.68
EDC + SSM	57.10	40.07	66.48	65.52	55.69	52.38	68.73	67.78
DAC + STZ	55.23	51.44	63.43	62.52	55.96	54.09	61.19	60.05
DAC + SSM	61.78	60.06	66.64	65.53	52.85	41.98	65.06	58.54
STZ + SSM	59.63	58.46	67.64	66.68	58.61	53.28	66.62	65.53
RLRs + CBRs								
W2V + EDC	75.13	75.07	60.24	60.58	78.55	78.50	90.77	89.72
W2V + DAC	77.70	77.70	78.26	78.19	85.33	85.05	89.77	89.72
W2V + STZ	72.41	71.89	74.62	74.39	84.15	83.79	91.84	91.52
W2V + SSM	76.15	76.15	**79.29**	**79.10**	82.57	82.24	**95.33**	**95.18**
WLM + EDC	66.89	66.79	77.68	76.82	80.37	79.49	95.20	94.11
WLM + DAC	60.91	60.55	66.13	65.75	85.05	84.71	85.98	85.98
WLM + STZ	76.02	72.61	75.35	75.14	83.18	83.02	95.05	95.05
WLM + SSM	65.33	65.28	66.89	66.99	64.96	63.01	87.31	85.98
XVC + EDC	63.64	63.47	67.85	66.28	86.03	85.98	94.39	94.14
XVC + DAC	65.16	65.01	69.63	68.32	86.02	85.05	91.59	91.59
XVC + STZ	65.28	65.14	71.52	70.89	87.30	86.92	92.52	92.20
XVC + SSM	63.65	63.53	69.88	68.13	69.31	65.14	93.01	92.52
HBT + EDC	67.02	66.51	69.18	69.04	85.56	85.05	88.35	87.85
HBT + DAC	64.14	63.99	68.23	68.47	82.62	82.24	90.14	88.79
HBT + STZ	68.03	68.03	67.29	67.22	83.18	82.92	86.17	85.98
HBT + SSM	67.02	66.82	68.83	68.64	65.96	63.21	89.31	88.79

Table 4.2: Performance results for models trained using different combinations of RLRs and CBRs. All scores are reported as percentages and represent the mean across five folds

resentations, suggesting complementary behavior. CBR+CBR fusions remain weaker, reflecting their lower individual performance. Notably, certain RLR+CBR pairs yield greater gains than RLR+RLR, demonstrating that combining RLRs' rich prosodic cues with CBRs' fine-grained acoustic features offers stronger complementarity. Both the baseline concatenation fusion and **HYFuse** reveal complementary behavior between

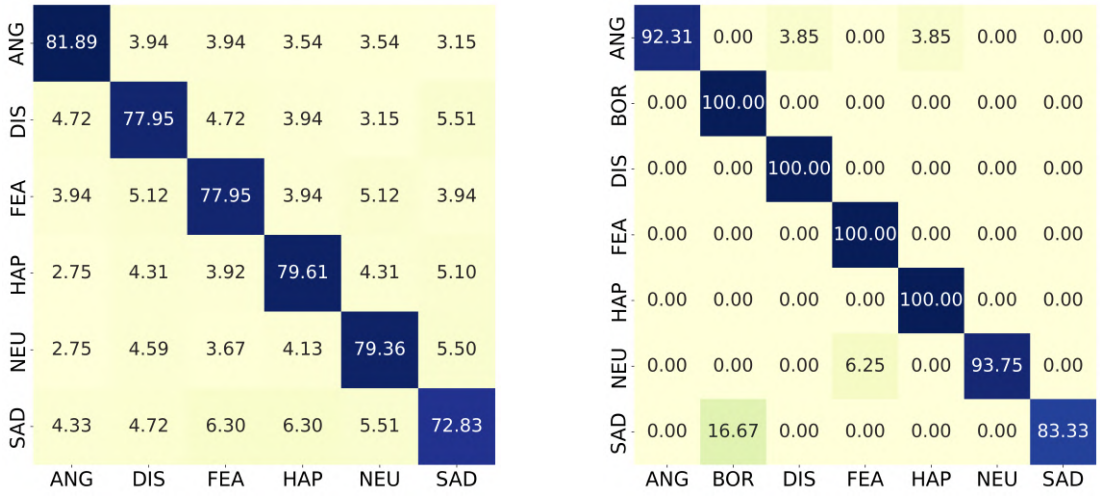


(a)



(b)

Figure 4.2: t-SNE visualizations: (a) CNN (HuBERT) (b) **`HYFuse (Wav2vec2 + Soundstream)`**



(a) (b)

Figure 4.3: Confusion matrices: (a) CREMA-D, (b) Emo-DB; y-axis denotes the True Values and the x-axis denotes the Predicted Values.

RLRs and CBRs, but **`HYFuse`** captures it more effectively by aligning features in hyperbolic space, preserving hierarchical relationships and reducing distortion. The

Wav2vec2–SoundStream pair achieves the highest accuracy, outperforming all individual PTM representations, homogeneous fusions, and baseline concatenation, thus setting new SOTA results as the individual RLRs were already SOTA for SER wen Yang *et al.* (2021). *This supports our hypothesis that heterogeneous RLR+CBR fusion is most effective for SER*. While not all RLR+CBR pairs surpass RLR+RLR fusions, strong-performing CBRs combined with strong RLRs yield notable gains. For example, WavLM+EnCodec performs competitively, though slightly below the top combination, indicating that improvements depend on the specific characteristics of the paired representations. Overall, selecting optimal RLR–CBR pairs is crucial, and **HYFuse** effectively exploits their complementary strengths. Figure 4.2 provides t-SNE plots on CREMA-D contrasting HuBERT—the strongest individual PTM—with **HYFuse** using Wav2vec2+SoundStream. **HYFuse** displays more distinct and well-separated emotional clusters and thus augmenting our findings. Additionally, we show the confusion matrices for the Wav2vec2+SoundStream fusion on both datasets are shown in Figure 4.3.

Statistical Significance: To evaluate the reliability of the observed performance improvements, paired McNemar's tests were conducted on identical test sets. The proposed **HYFuse** framework (Wav2vec2 + SoundStream) was compared against the strongest concatentation-based fusion baseline utilizing the same representations. The resulting p-values were 0.00043 for CREMA-D and 0.00766 for Emo-DB. As all p-values are substantially below the commonly adopted significance threshold of 0.05, the improvements achieved by **HYFuse** can be regarded as statistically significant across all evaluated datasets.

4.2 MATA

In this section, we outline the motivation behind exploring multimodal PTMs for and their fusion for improved NVER. We then introduce our proposed framework, **MATA** Phukan *et al.* (2025*a*), which employs manifold-to-manifold alignment via optimal transport to achieve enhanced NVER performance.

4.2.1 Background and Motivation

Emotion recognition research has primarily focused on verbal speech, leveraging both handcrafted spectral features Kishore and Satish (2013) and, more recently, speech PTMs Chen and Rudnicky (2021). Speech PTMs such as WavLM Diatlova *et al.* (2024), Wav2vec2 Pepino *et al.* (2021), and HuBERT Morais *et al.* (2022) have proven effective for capturing emotional cues, typically through fine-tuning or as feature extractors for downstream tasks. However, non-verbal human vocalizations remain underexplored, with only a few notable works Hsu *et al.* (2021*a*); Tzirakis *et al.* (2023); Xin *et al.* (2024). Non-verbal human vocalizations—such as laughter, cries, and sighs—convey rich emotional information that enhances everyday communication. Emotion recognition from these cues has wide-ranging applications in healthcare, human–computer interaction, customer service, and security. In this work, we specifically focus on NVER. Moreover, multimodal PTMs—despite their potential for richer emotional interpretation—have seen no application in NVER. *In this work, we address this gap by evaluating multimodal PTMs for NVER, hypothesizing that their multimodal pre-training enhances contextual understanding, enabling better discrimination of subtle and ambiguous emotional cues compared to speech PTMs.* We test this by comparing SOTA multimodal PTMs (LB, IB) with leading speech PTMs (WavLM, Unispeech-SAT, Wav2vec2) using their extracted representations with a simple CNN classifier on benchmark NVER datasets (ASVP-ESD, JNV, VIVAE).

Motivated by prior research in areas such as speech recognition Arunkumar *et al.* (2022) and our work in SSD Phukan *et al.* (2024*a*), which has shown that combining PTMs can yield superior results, we present the first exploration of PTM representations fusion for enhanced NVER. To this end, we introduce **MATA** (Intra-**M**odality **A**lignment through **T**ransport **A**ttention), a novel framework that leverages optimal transport as fusion mechanism for effectively combining PTMs representations. In this section, we first determine the strongest PTM representations for NVER and then develop fusion strategies around them. Our work also highlights that NVER plays a vital role in promoting inclusivity in affective and secure VIS by enabling emotion recognition for minimal speaking individuals, such as those with autism or other communication disorders, thereby broadening accessibility and user support.

The main contributions are:

- A comprehensive investigation for demonstrating the effectiveness of multimodal PTMs for NVER, achieving superior performance over unimodal speech PTMs.
- A novel **MATA** fusion framework that surpasses individual PTMs and baseline fusion methods, delivering SOTA across all evaluated datasets.
- To the best of our knowledge, we are the first to explore fusion of PTM representations for improved NVER.

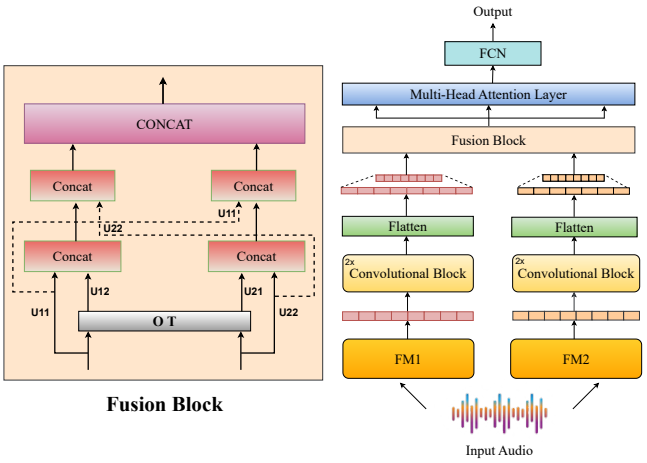


Figure 4.4: **MATA**: FM1 and FM2 denote PTM 1 and PTM 2, respectively. U11 and U22 correspond to features extracted from the individual PTM branches, whereas U12 and U21 represent features transferred from FM2 to FM1 and from FM1 to FM2, respectively

4.2.2 Modeling

We first describe the downstream modeling with individual PTMs representations followed by the proposed fusion framework, **MATA**.

Individual PTMs representations: For each PTM, extracted representations pass through two 1D-CNN blocks consisting of 1D-CNN layer followed by maxpooling: the first with 64 filters (kernel size 3), the second with 128 (kernel size 3). Features are flattened, fed to a 128-unit dense layer, and classified via a softmax output (units = #emotion classes).

MATA: Framework: The architecture of **MATA** is illustrated in Figure 4.4, where OT is employed as the core mechanism for fusion of PTM representations. OT aligns the underlying geometric structure of heterogeneous PTM representation manifolds. Although optimal transport is widely recognized as a SOTA technique in multimodal fusion Pramanick *et al.* (2021), its use for aligning PTM representations in NVER—and

more broadly within voice-based affective computing—has not yet been explored. OT provides a principled means of transferring the distributional geometry of one representational space onto another, enabling manifold-to-manifold alignment across PTMs. Firstly, each PTM's extracted representations through two 1D-CNN blocks (32 and 64 filters, kernel size 3), followed by flattening and linear projection to 120 dimensions for computational efficiency. For fusion, we compute an OT distance matrix D between feature matrices f_a and f_b from two PTMs using normalized euclidean distance:

$$D = \frac{\|f_a - f_b\|_2}{\max(\|f_a - f_b\|_2)}$$

The Sinkhorn algorithm produces the OT plan $\Gamma = \text{Sinkhorn}(D)$. Transported features are obtained as: $f_b \rightarrow f_a = \Gamma f_b, \quad f_a \rightarrow f_b = \Gamma^\top f_a$. These transported features are concatenated with the corresponding original features to form fused_a and fused_b. Following this, the aligned features are again combined with the original features from the opposite PTM feature space. The resulting representations are concatenated once more and subsequently fed into the Multi-Head Attention (MHA) module block (8 heads) for enhanced cross-feature interaction:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Here, Q, K, and V are query, key, value matrices and derived from the final fused features. Depending on the PTM pair, **MATA** contains 4.0M–4.5M trainable parameters.

4.2.3 Experiments

Datasets: We use ASVP-ESD, JNV and VIVAE in our experiments. Detailed information about the datasets is given in Chapter 3.

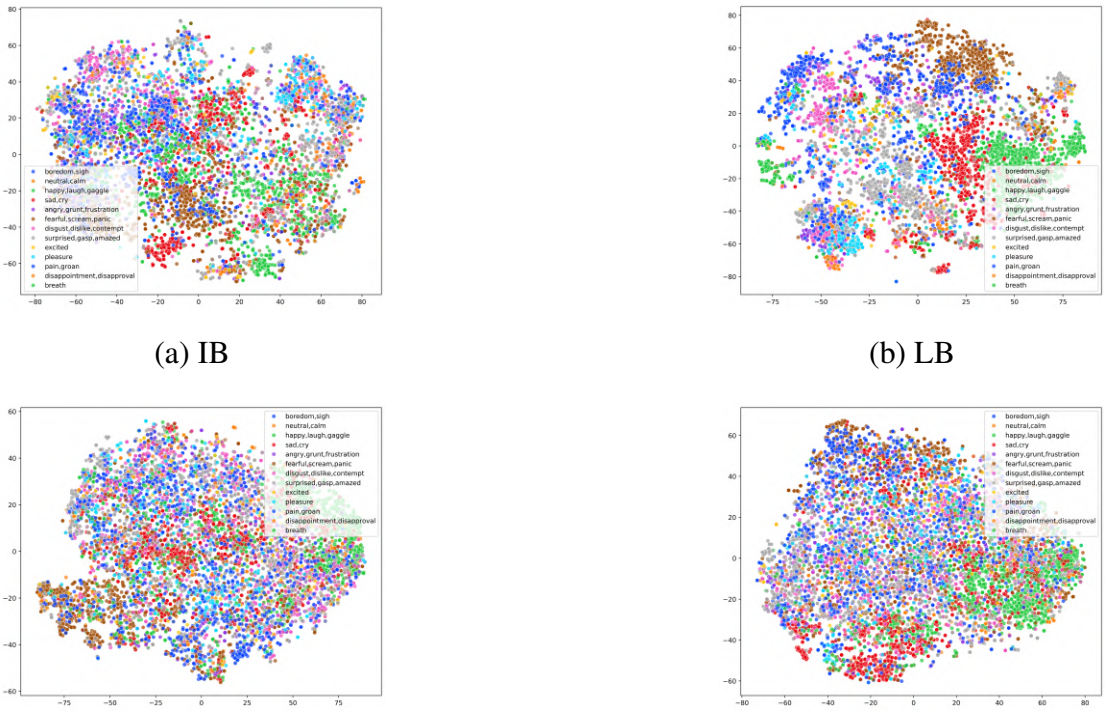
Pre-Trained Models: We use LB, IB as multimodal PTMs and Wav2vec2, WavLM (Base), Unispeech-SAT as speech PTMs. More details are given in Chapter 2.

Training Details: All models are trained for 50 epochs using the Adam optimizer with a learning rate of 1e-3 and a batch size of 32. Cross-entropy is used as the loss function, while dropout is employed to mitigate overfitting.

Table 4.3: Evaluation scores (%) averaged over five folds. Abbreviations: UNI: Unispeech-SAT, WLM: WavLM, W2V: Wav2vec2. F1-Score is macro-average F1 score.

Features	ASVP_ESD		JNV		VIVAE		CREMA-D	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Individual Representations								
LB	75.55	67.55	73.65	72.51	69.12	68.83	63.67	63.26
IB	62.03	49.11	63.10	60.97	55.30	54.63	63.67	63.68
UNI	49.90	35.86	57.14	53.81	36.87	35.78	63.26	63.20
WLM	46.98	32.77	62.07	58.33	35.94	34.39	55.88	55.92
W2V	60.04	48.51	57.14	59.27	46.54	45.65	59.84	59.82
Fusion with Concatenation								
LB+IB	**76.34**	64.58	72.62	72.26	67.48	67.28	**71.46**	**71.71**
LB+UNI	74.09	65.29	67.86	66.36	61.75	61.70	68.06	67.70
LB+WLM	72.90	64.80	70.24	70.92	65.44	65.35	66.35	65.96
LB+W2V	73.76	62.20	70.24	69.18	60.83	60.57	69.04	68.93
IB+UNI	65.74	55.45	58.33	52.25	53.00	52.42	72.60	72.66
IB+WLM	66.14	56.83	58.33	52.25	58.53	57.28	69.31	69.28
IB+W2V	67.20	55.53	57.14	55.19	56.22	55.73	68.57	68.67
UNI+WLM	53.61	38.98	44.05	39.10	44.70	43.88	66.76	66.69
UNI+W2V	60.70	49.16	47.66	46.18	49.31	49.51	68.84	68.81
WLM+W2V	59.64	45.66	51.19	41.98	48.85	46.97	68.10	68.16
Fusion with OT								
LB+IB	**76.41**	**68.79**	**77.03**	**76.14**	**70.05**	**69.80**	62.12	62.05
LB+UNI	75.61	67.52	70.24	71.90	61.29	60.33	59.44	58.84
LB+WLM	75.48	66.97	69.05	70.48	62.21	61.30	58.16	58.14
LB+W2V	75.48	67.75	67.86	66.44	66.82	66.50	58.03	57.84
IB+UNI	64.88	53.19	64.29	62.59	57.14	56.49	62.46	62.24
IB+WLM	64.68	54.40	59.52	59.05	57.14	56.65	59.17	59.06
IB+W2V	67.00	55.41	61.90	61.21	59.91	59.30	57.76	57.62
UNI+WLM	55.47	45.62	59.52	56.60	41.01	39.40	60.51	60.45
UNI+W2V	60.77	48.43	47.62	47.54	46.54	45.12	58.63	58.68
WLM+W2V	60.64	51.85	60.71	60.60	49.77	49.04	58.97	59.09
Fusion with MATA								
LB+IB	**76.47**	**70.35**	**77.40**	**76.19**	**75.12**	**74.63**	**72.64**	**72.62**
LB+UNI	75.41	66.51	71.43	72.17	65.44	64.94	69.85	69.98
LB+WLM	75.75	**70.25**	73.81	**73.08**	69.12	68.64	66.29	66.30
LB+W2V	75.81	68.49	**75.00**	72.39	**69.59**	**69.15**	66.55	66.66
IB+UNI	67.93	60.15	61.90	62.17	60.37	59.43	**71.32**	**71.49**
IB+WLM	65.94	57.44	61.90	60.15	61.75	61.12	68.64	68.72
IB+W2V	68.06	61.14	64.29	64.74	60.37	60.13	68.57	68.57
UNI+WLM	56.66	46.55	46.43	43.78	45.16	43.99	66.82	66.88
UNI+W2V	61.43	52.89	53.57	55.02	48.85	47.58	70.99	71.06
WLM+W2V	62.36	52.29	59.63	57.71	50.69	49.24	67.36	67.49

Experimental Results and Discussion: Table 4.3 summarizes our findings. Among individual PTMs, LB consistently achieved the highest scores across all NVER datasets, reaching 75.55% accuracy / 67.55% F1 on ASVP-ESD, 73.65% / 72.51% on JNV, and 70.05% / 69.80% on VIVAE, outperforming all speech PTMs. *These results confirm our hypothesis that multimodal PTMs, through multimodal pre-training, capture subtle emotional cues often missed by speech PTMs.* t-SNE visualizations (Figure 4.5) further show clearer emotion clusters for multimodal PTMs. Fusing multimodal PTM representations with **MATA** achieved the strongest overall results, outperforming both individual PTMs and a concatenation-based fusion baseline. The baseline mirrors the



(a) IB

(b) LB

(c) WavLM

(d) Unispeech-SAT

Figure 4.5: t-SNE visualizations of raw representations extracted from PTMs on the ASVP-ESD

MATA architecture but replaces the OT fusion block with simple concatenation. The performance gap highlights not only the complementary nature of multimodal PTMs but also the effectiveness of **MATA**'s OT-driven fusion. Interestingly, combinations of multimodal PTMs with speech PTMs also surpassed individual PTMs performance and consistently outperformed the concatenation baseline. An ablation without the MHA block (Table 4.3, "Fusion with OT") still exceeded individual PTMs results and, in couple of cases, matched **MATA** performance.

Statistical Significance: The statistical significance of the observed performance gains was assessed using paired McNemar's tests on identical test sets. Pairwise comparisons between **MATA** (LB + IB) and the strongest OT-based fusion baseline employing the same PTMs yielded p-values of 0.00672, 0.00471, and 0.00015 for ASVP_ESD, JNV, and VIVAE, respectively. All obtained p-values are below the conventional significance threshold of 0.05, indicating that the improvements achieved by **MATA** are statistically significant across all evaluation settings.

Additional Experiments: To assess the generalizability of **MATA**, we evaluated it on CREMA-D which is an benchmark SER dataset. As shown in Table 4.3, multimodal

PTMs consistently outperformed speech PTMs. Fusion of multimodal PTMs, LB and IB via **MATA** achieves the best results, further confirming the framework's effectiveness.

4.3 MERLINN

In this section, we first present the background and motivation for SSD in the context of children's voices. We then demonstrate that existing SOTA SSD approaches shows significant performance degradation when exposed to synthetic children speech (Figure 4.6). Motivated by this limitation, we introduce a novel and dedicated task for detecting synthetic children speech, along with a newly curated dataset to support this problem setting. We further show that multimodal PTMs are particularly effective for SSD in this challenging scenario. Finally, we advance to the introduction of **MERLINN**, a Geometry Guided Alignment framework for PTM representations, which leverages geometric alignment principles to enhance detection performance.

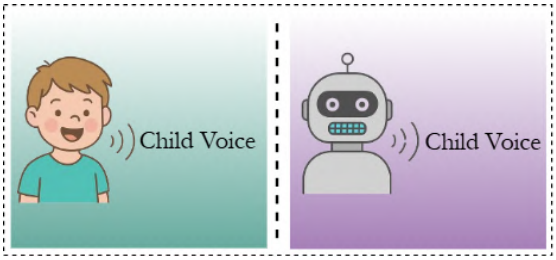


Figure 4.6: Synthetic Children Voice

4.3.1 Background and Motivation

Beyond vocoder-based synthesis (These type of vocoder-based synthesis was the focus in our 2024 SSD work Chapter 5 **MiO** Phukan *et al.* (2024*a*)), a new class of synthetic speech has emerged alongside the rapid progress of LALMs Lu *et al.* (2024). While earlier speech synthesis pipelines predominantly relied on vocoders, contemporary LALMs integrate NACs such as EnCodec Défossez *et al.* (2023), SoundStream Zeghidour *et al.* (2021), and DAC Kumar *et al.* (2024) as core components, enabling high-fidelity and expressive speech generation. However, existing SSD models—largely trained on vocoder-generated synthetic speech—struggle to generalize to such scenarios due to substantial distributional shifts in the synthesized speech. This limitation

has prompted the development of new benchmarks and detection strategies specifically designed for NAC-generated speech Xie *et al.* (2025*a*), signaling a significant shift in the SSD landscape. Despite these advances, current NAC-generated speech detection datasets predominantly consist of adult speech, which may restrict the applicability of these methods to other demographic groups. As noted by Kulkarni *et al.* (2025), children's speech exhibits pronounced acoustic and prosodic differences compared to adult speech. Such distinctions introduce additional challenges and position children as a particularly vulnerable demographic within the evolving NAC-generated speech threat landscape.

To address this gap, we introduce the CodecFake-Kids (CFK) dataset and formulate a new task termed Children CodecFake Detection (CCFD), which we use throughout this section for convenience. The CFK dataset comprises both bonafide and NAC-synthesized child speech, covering a wide range of age groups, languages, and NAC types. We benchmark SOTA NAC-generated speech detection models—such as AASIST Jung *et al.* (2022)—trained on existing adult-centric NAC-generated speech datasets and observe a pronounced degradation in performance when evaluated on CFK. These findings underscore the limitations of current approaches and highlight the need for child-specific NAC-generated speech detection frameworks. Motivated by this observation, *we hypothesize that multimodal PTMs, including LB and IB, are particularly well suited for CCFD. This hypothesis is grounded in their cross-modal pretraining, which may implicitly encode child-related contextual information through exposure to visual modalities such as faces and scenes involving children*. To validate this claim, we conduct a comprehensive comparative evaluation of SOTA speech PTMs and multimodal PTMs, with extensive experiments on the CFK dataset providing strong empirical support for our hypothesis.

Furthermore, inspired by prior successes in PTM fusion for speech processing tasks, such as our work on SSD (Interaction Guided Alignment **MiO** Phukan *et al.* (2024*b*)), and speech recognition Arunkumar *et al.* (2022)—we investigate representation fusion as a means to further enhance CCFD performance. To this end, we propose **MERLINN** (**M**ulti-geometry Cent**ER**ed Kernel A**LI**g**N**me**N**t), a novel framework for effective PTM fusion. **MERLINN** aligns latent representations using CKA in both euclidean and hyperbolic spaces. We introduce a novel formulation of CKA in hyperbolic space to accommodate the geometric heterogeneity and potential hierarchical structure inher-

ent in PTM representations. By leveraging the complementary properties of flat (euclidean) and curved (hyperbolic) geometries, **MERLINN** enables effective alignment across geometrically diverse latent spaces. Extensive experimental results demonstrate that **MERLINN** with fusion of LB and IB consistently achieves the best performance on the CCFD task, outperforming individual PTMs as well as strong baseline fusion methods.

The principal contributions are outlined as follows:

- We introduce the CFK dataset and formally define the novel task of CCFD, addressing a critical gap in the SSD literature.
- We evaluate SOTA SSD detectors trained on previous NAC-generated speech datasets on the CFK dataset and show that they generalize poorly to CCFD, underscoring the need for child-specific SSD frameworks.
- We hypothesize and empirically validate that multimodal PTMs (e.g., LB and IB), due to their cross-modal pre-training, are better suited for CCFD. A comprehensive comparison against SOTA speech PTMs proves this claim.
- We introduce **MERLINN**, a novel fusion framework that jointly aligns PTM representation spaces using CKA in both euclidean and hyperbolic geometries, achieving superior performance over individual PTMs and baseline fusion methods.

4.3.2 CodeFake-Kids Dataset

In this subsection, we present a comprehensive overview of the datasets used to construct the CFK dataset, the NACs employed for speech synthesis, and the end-to-end pipeline adopted for generating synthetic speech samples. All child speech data utilized in this work originate from publicly available datasets that were collected under appropriate consent and ethical oversight. No new recordings were collected, and no personal or identifiable information was used.

Children Speech Datasets: We select openly accessible benchmark datasets of children's speech, as the majority of children's speech corpora are either difficult to access or restricted due to privacy and ethical concerns Sobti *et al.* (2024); Kulkarni *et al.* (2025). Details of the selected datasets are provided below. **(i) Non-Native Children's English Speech (NNCES) (C0)** Radha *et al.* (2024) : This corpus comprises English utterances spoken by Indian children and was originally developed for children's speaker recognition. It includes speech from 50 children (25 female and 25 male), aged 8 to

12 years, all of whom are native Telugu speakers and use English as the medium of instruction in school. The dataset contains a total of 10000 utterances, evenly split between read and spontaneous speech, amounting to approximately 19 hours of audio recordings. **(ii) ChildMandarin (C1)** Zhou *et al.* (2024): This dataset is tailored for Mandarin-speaking children aged 3 to 5 years, an underrepresented age group in speech research. It comprises 41.25 hours of high-quality conversational speech collected from 397 children, with balanced gender representation, spanning 22 provinces across China. The corpus contains 40913 utterances, each with an average duration of 3.5 seconds. Designed to support robust evaluation, this dataset facilitates benchmarking of speech recognition and speaker verification models on young children's speech. **(iii) Samromur (C2)** Mena *et al.* (2022): This corpus represents the first large-scale Icelandic speech dataset exclusively focused on children's voices. It contains speech recordings from children aged 4 to 17 years, comprising 131 hours of read speech across 137597 utterances from 3175 unique speakers. The dataset is partitioned into training (127 hours), development, and test splits, with the test set carefully balanced to ensure diversity across age and gender. All audio samples are recorded at a sampling rate of 16 kHz.

Neural Audio Codecs: We adopt the NACs employed by Lu *et al.* (2024) and Wu *et al.* (2024*b*), selecting SOTA, publicly available codecs that are widely accessible and reproducible. The NACs considered in this study include DAC, EnCodec, SoundStream, SpeechTokenizer, FunCodec, AudioDec, and SNAC. Accounting for multiple versions of these codecs, the total number of NAC configurations used amounts to 12. Detailed descriptions of the selected NACs are provided in Chapter 2.

Generation Pipeline of CFK: To construct the CFK dataset, we adopt a controlled synthesis pipeline inspired by prior work on NAC-generated speech generation Wu *et al.* (2024*b*). The process involves re-synthesizing bona fide children's speech using a diverse set of NACs that underpin modern LALMs, such as AudioLM, which is built on SoundStream Borsos *et al.* (2023); Lu *et al.* (2024). We begin with the three publicly available children's speech corpora described earlier, treating each original utterance as a bonafide reference sample. To generate NAC-based synthetic samples, we employ a codec synthesis–resynthesis pipeline. Each original waveform is first encoded into a discrete latent representation using the pre-trained encoder of a given NAC. The latent codes are then decoded back into the waveform domain using the correspond-

ing decoder, producing a synthetic (fake) utterance. While this process preserves the linguistic content and speaker identity, it introduces subtle NAC-induced artifacts, resulting in realistic yet synthetic child speech samples. This procedure is applied across 12 NAC configurations, yielding parallel versions of the dataset in which each bonafide utterance has a one-to-one corresponding synthetic counterpart for every codec. For the NNCES dataset Radha *et al.* (2024), which includes both read and spontaneous speech subsets, synthetic samples are generated independently for each subset. For the ChildMandarin Zhou *et al.* (2024) and Samromur Mena *et al.* (2022) datasets, we adhere to the official train, validation, and test splits provided, applying codec-based synthesis within each partition. We define two evaluation settings: seen and unseen. In the seen setting, the test set includes synthetic utterances generated using the same NACs encountered during training, including variants of SNAC, DAC, as well as EnCodec, SoundStream, and SpeechTokenizer. In contrast, the unseen setting evaluates generalization by testing on samples synthesized using previously unseen codecs, such as FunCodec and AudioDec variants. Together, datasets C0, C1, and C2 collectively constitute the CFK dataset.

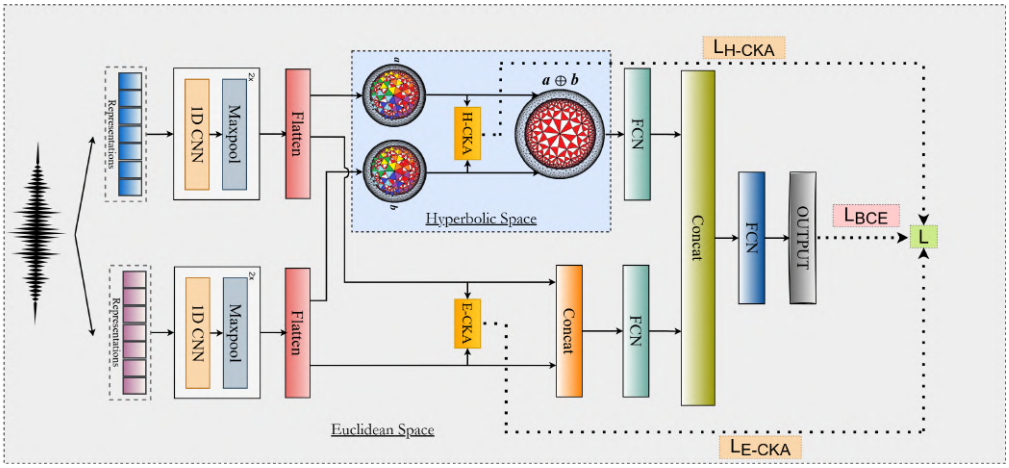


Figure 4.7: **MERLINN**; $x \oplus y$ stands for mobius addition; L, L_{BCE}, $L_{\text{H-CKA}}$, $L_{\text{E-CKA}}$ denotes the total loss, binary cross-entropy loss, Hyperbolic CKA loss, and Euclidean CKA loss respectively

4.3.3 Methodology

In this section, we present detailed descriptions of the PTMs considered in this study, followed by the modeling approaches using individual PTMs, and finally introduce the proposed fusion framework, **MERLINN**.

Pre-Trained Models: We select SOTA speech PTMs, including WavLM (Base), Unispeech-SAT, Wav2vec2, Whisper, XLS-R (300M), and MMS. For multimodal PTMs, we employ LB and IB. Additional details regarding these PTMs are provided in Chapter 2.

Individual PTM Representations Modeling: As downstream models for PTMs, we adopt two widely used architectures: CNN used in our SSD work Phukan *et al.* (2024*a*) and AASIST Ba *et al.* (2023); Wu *et al.* (2024*b*); Lu *et al.* (2024), following established practices in SSD. For AASIST, we utilize the original architecture proposed in Jung *et al.* (2022). The CNN architecture comprises two sequential 1D convolutional layers with 64 and 128 filters, respectively, each using a kernel size of 3 and ReLU activation. Each convolutional block is followed by a max-pooling layer to downsample the feature maps. The resulting representations are then flattened and passed through a fully connected layer with 128 ReLU-activated neurons, followed by a sigmoid output layer for binary classification.

MERLINN: Framework

We introduce **MERLINN** (Figure 4.7), a novel framework for the effective fusion of PTM representations. The overall architecture of **MERLINN** is illustrated in Figure 1. **MERLINN** aligns latent representations extracted from heterogeneous PTMs using CKA in both euclidean and hyperbolic spaces. CKA is commonly used to quantify representational similarity between embedding spaces, where higher CKA values indicate stronger alignment Kornblith *et al.* (2019), and has also been employed as a fusion mechanism Yun *et al.* (2022). We extend CKA to hyperbolic space to capture geometric discrepancies and exploit the hierarchical structure often embedded in PTM representations. By jointly leveraging the complementary curvature properties of euclidean (flat) and hyperbolic (negatively curved) geometries, **MERLINN** enables robust alignment across diverse PTMs. Representations from each PTM are first processed by a CNN identical to the individual PTM modeling setup described earlier. The resulting features are flattened and passed to the CKA modules. All CKA computations are performed in a batch-wise manner.

Euclidean CKA (E-CKA): Let

$$\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]^\top \in \mathbb{R}^{m \times d}, \quad \mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_m]^\top \in \mathbb{R}^{m \times d},$$

be batch-wise representations from two PTMs branch after flattening, where m is the batch size and d is the representation dimension. The linear kernel matrices are

$$\mathbf{K}_U = \mathbf{U}\mathbf{U}^\top, \quad \mathbf{K}_V = \mathbf{V}\mathbf{V}^\top.$$

Kernel centering is performed using

$$\mathbf{C} = \mathbf{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top,$$

yielding

$$\widehat{\mathbf{K}}_U = \mathbf{C}\mathbf{K}_U\mathbf{C}, \quad \widehat{\mathbf{K}}_V = \mathbf{C}\mathbf{K}_V\mathbf{C}.$$

Euclidean CKA is then computed as

$$\mathrm{CKA}_{\mathrm{E}} = \frac{\langle \widehat{\mathbf{K}}_U, \widehat{\mathbf{K}}_V \rangle_F}{\|\widehat{\mathbf{K}}_U\|_F \cdot \|\widehat{\mathbf{K}}_V\|_F},$$

where $\langle \cdot, \cdot \rangle_F$ and $\| \cdot \|_F$ denote the Frobenius inner product and norm, respectively.

Hyperbolic CKA (H-CKA): Euclidean representations are mapped to the Poincare ball $\mathbb{D}^d = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\| < 1\}$ using the exponential map:

$$\mathbf{P} = \left[\tanh(\|\mathbf{u}_i\|)\frac{\mathbf{u}_i}{\|\mathbf{u}_i\|}\right]_{i=1}^{m}, \quad \mathbf{Q} = \left[\tanh(\|\mathbf{v}_i\|)\frac{\mathbf{v}_i}{\|\mathbf{v}_i\|}\right]_{i=1}^{m}.$$

The hyperbolic distance is given by

$$d_{\mathbb{D}}(\mathbf{p}_i, \mathbf{q}_j) = \cosh^{-1}\left(1 + 2\frac{\|\mathbf{p}_i - \mathbf{q}_j\|^2}{(1 - \|\mathbf{p}_i\|^2)(1 - \|\mathbf{q}_j\|^2)}\right).$$

Using this distance, RBF kernel matrices are constructed as

$$\mathbf{K}_{ij}^{(\mathbb{D})} = \exp\left(-\frac{d_{\mathbb{D}}^2(\mathbf{p}_i, \mathbf{p}_j)}{2\tau^2}\right), \quad \mathbf{L}_{ij}^{(\mathbb{D})} = \exp\left(-\frac{d_{\mathbb{D}}^2(\mathbf{q}_i, \mathbf{q}_j)}{2\tau^2}\right),$$

where τ is the kernel bandwidth. Centering yields

$$\widehat{\mathbf{K}}^{(\mathbb{D})} = \mathbf{C}\mathbf{K}^{(\mathbb{D})}\mathbf{C}, \quad \widehat{\mathbf{L}}^{(\mathbb{D})} = \mathbf{C}\mathbf{L}^{(\mathbb{D})}\mathbf{C}.$$

Hyperbolic CKA is computed as

$$\mathrm{CKA}_{\mathrm{H}} = \frac{\langle \widehat{\mathbf{K}}^{(\mathbb{D})}, \widehat{\mathbf{L}}^{(\mathbb{D})} \rangle_F}{\|\widehat{\mathbf{K}}^{(\mathbb{D})}\|_F \cdot \|\widehat{\mathbf{L}}^{(\mathbb{D})}\|_F}.$$

Further, the hyperbolic representations are fused using mobius addition,

$$\mathbf{r}_i^{(\mathbb{D})} = \mathbf{p}_i \oplus \mathbf{q}_i,$$

and projected back to euclidean space via the logarithmic map:

$$\mathbf{r}_i^{(\text{fused})} = \tanh^{-1}(\|\mathbf{r}_i^{(\mathbb{D})}\|) \frac{\mathbf{r}_i^{(\mathbb{D})}}{\|\mathbf{r}_i^{(\mathbb{D})}\|}.$$

In parallel, euclidean representations are concatenated after E-CKA computation. The outputs of the E-CKA and H-CKA branches are each passed through a fully connected layer with 90 neurons, concatenated, and then processed by a dense layer with 60 neurons, followed by a sigmoid output layer for binary classification. The total training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} - \eta \left(\mathrm{CKA}_{\mathrm{E}} + \mathrm{CKA}_{\mathrm{H}} \right),$$

where the negative sign encourages maximization of CKA and η controls the trade-off between classification accuracy and representation alignment. The number of trainable parameters in **MERLINN** ranges from 4.2M to 5.5M, depending on the PTM representation dimensionality.

4.3.4 Training Pipeline

We use Adam optimizer with binary cross-entropy loss for training. All models are trained for 50 epochs using a batch size of 32 and a learning rate of 1e-3. To reduce overfitting, early stopping and dropout are applied during training. For experiments involving **MERLINN**, the weighting parameter λ is fixed at 0.4, as determined through preliminary experimentation.

PTM	C0			C1			C2			C0			C1			C2		
	A	F1	EER	A	F1	EER	A	F1	EER	A	F1	EER	A	F1	EER	A	F1	EER
	AASIST									CNN								
W2V	84.52	78.87	13.24	85.30	80.06	16.71	89.95	86.96	17.88	86.75	81.48	17.21	87.15	82.12	14.94	85.62	80.28	18.37
WLM	86.10	81.08	14.91	87.42	82.19	15.23	**91.13**	**87.97**	17.91	89.42	84.11	13.65	89.97	84.87	13.81	90.21	85.13	15.33
UNI	83.42	77.47	15.36	82.95	76.83	19.93	85.55	80.22	20.12	85.26	79.27	17.58	84.52	78.36	16.82	84.98	78.83	17.25
XLR	88.45	84.93	**10.52**	86.93	82.43	14.91	90.50	87.17	**13.03**	91.10	88.17	**9.24**	90.42	87.48	11.63	91.55	88.64	**12.45**
WSR	85.78	80.36	11.70	86.34	81.03	16.57	86.12	84.38	18.42	87.61	82.08	12.22	88.29	83.06	13.88	86.32	80.92	17.73
MMS	**89.12**	**85.69**	11.30	**88.94**	**85.38**	**13.63**	89.78	86.17	16.84	**91.43**	**88.66**	11.25	**91.15**	**88.32**	**10.72**	**91.78**	**89.02**	13.91
LB	**90.12**	**86.42**	**6.96**	**90.87**	**88.03**	**7.47**	**91.89**	**89.92**	**12.54**	**93.14**	**90.04**	**6.29**	**91.87**	**87.98**	**7.41**	**92.92**	**90.21**	**8.11**
IB	**91.31**	**88.06**	**5.83**	**89.55**	**85.88**	**10.18**	**93.72**	**90.02**	**13.51**	**92.23**	**88.28**	**5.57**	**92.98**	**90.13**	**6.19**	**92.00**	**88.18**	**9.26**

Table 4.4: Performance comparison across different PTMs and downstream architectures. Abbreviations: W2V—Wav2vec2, WLM—WavLM, UNI—Unispeech-SAT, XLR—XLS-R, WSR—Whisper, MMS—MMS, LB—LanguageBind, and IB—ImageBind. Evaluation metrics include Accuracy (A, %), macro F1-score (F1, %), and Equal Error Rate (EER, %). Datasets: C0—NNCES, C1—ChildMandarin, C2—Samromur. For each metric, the top three scores (column-wise) are highlighted in blue (best), green (second best), and yellow (third best). The same abbreviation set and highlighting scheme are used in Table 4.5.

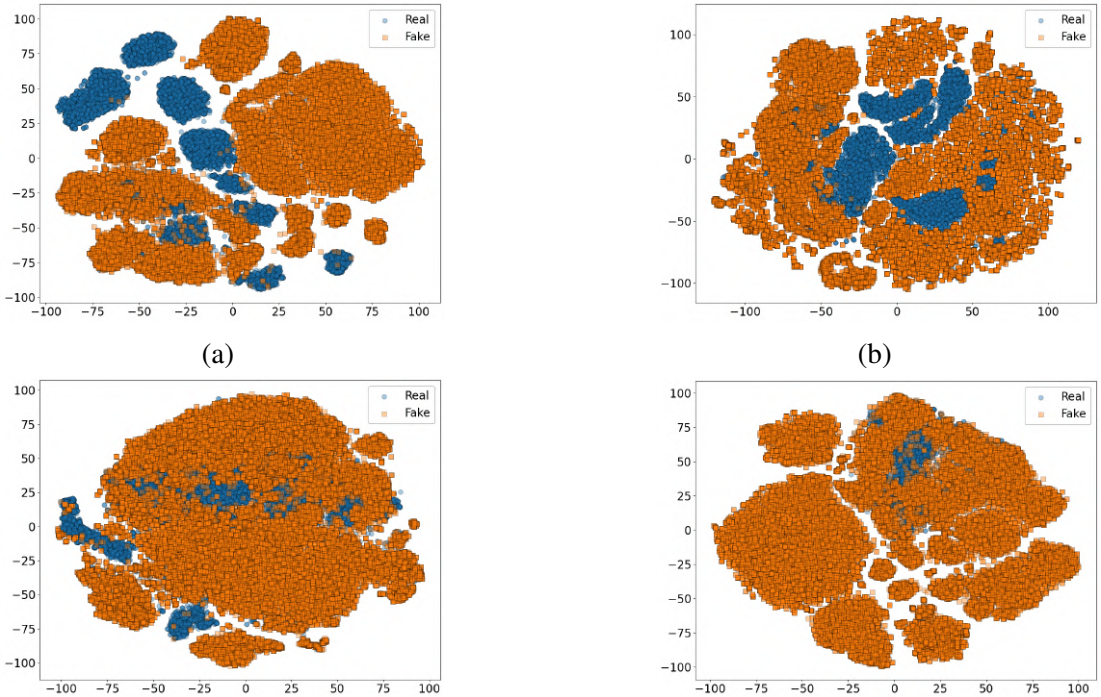


Figure 4.8: t-SNE Plots of raw PTM representations (a) LB (b) IB (c) MMS (d) WavLM

4.3.5 Experimental Results

We report experimental results across multiple evaluation settings. Performance is assessed using standard metrics commonly adopted in SSD research Jung *et al.* (2022); Ba *et al.* (2023); Phukan *et al.* (2024*b*). For C0, we alternate between the read and spontaneous splits for training and testing—training on one split while testing on the

other, and vice versa—with final scores averaged across both configurations. For C1 and C2, we adhere to the official train–test partitions, and results are reported on the corresponding test sets. Notably, evaluation is performed by jointly considering both seen and unseen test samples, and the reported scores reflect performance over the entire test set.

Pairs	Only E-CKA									MERLINN								
	C0			C1			C2			C0			C1			C2		
	A	F1	EER	A	F1	EER	A	F1	EER	A	F1	EER	A	F1	EER	A	F1	EER
W2V-WLM	85.88	83.62	12.95	90.96	88.12	16.27	90.40	89.12	14.84	88.52	87.49	10.40	93.18	91.44	12.92	93.29	92.20	12.00
W2V-UNI	84.05	78.33	15.34	87.81	85.07	17.56	88.71	86.15	18.88	86.08	84.86	12.33	89.84	88.64	14.26	91.25	90.03	15.70
W2V-XLR	90.40	87.53	9.75	87.59	85.91	14.22	91.59	90.37	12.63	93.21	92.20	6.14	90.29	88.95	11.91	94.55	93.21	10.44
W2V-WSR	86.80	80.94	13.07	89.15	86.85	15.98	90.36	88.01	16.19	88.90	87.05	9.86	91.96	90.23	12.91	93.33	91.95	13.09
W2V-MMS	88.94	84.62	10.50	90.94	87.38	13.26	87.91	86.16	16.92	91.77	90.15	6.78	93.52	91.82	11.17	90.14	88.85	14.76
W2V-LB	**92.80**	**89.62**	6.54	89.75	87.12	9.35	92.89	90.92	12.26	**95.03**	**93.93**	3.98	92.39	91.03	6.61	95.10	93.83	8.39
W2V-IB	92.10	**90.43**	6.10	91.98	89.95	9.35	93.56	91.19	12.89	94.75	93.14	3.76	94.71	93.55	6.59	96.55	94.91	9.78
WLM-UNI	84.76	79.27	13.14	89.18	86.08	18.94	90.84	88.09	16.42	87.44	85.60	9.59	91.41	90.38	16.31	93.11	91.90	12.53
WLM-XLR	87.59	83.55	11.45	89.51	87.75	13.69	92.81	88.76	16.27	90.47	89.16	8.14	91.91	90.00	10.77	95.07	93.82	13.15
WLM-WSR	87.28	82.05	12.07	89.18	86.60	14.32	90.52	88.69	16.35	89.54	87.96	8.27	91.58	90.36	10.32	93.03	91.94	14.26
WLM-MMS	89.38	85.75	10.72	91.94	90.80	12.44	90.50	89.17	15.33	91.91	90.27	8.11	94.96	92.68	9.25	92.97	91.35	12.12
WLM-LB	90.59	88.85	6.13	92.87	90.61	7.07	93.89	89.85	13.11	93.98	92.17	2.74	**95.48**	93.31	4.30	**96.89**	94.26	10.32
WLM-IB	90.18	89.50	**5.95**	91.37	90.02	8.98	94.05	90.11	13.82	93.02	91.44	**2.09**	94.68	92.78	5.81	**96.89**	**95.13**	10.12
UNI-XLR	85.93	80.95	14.00	81.94	79.86	16.88	89.53	88.20	14.44	88.93	87.60	11.12	84.94	83.36	14.13	92.17	91.30	10.09
UNI-WSR	84.83	79.70	13.38	80.15	78.12	16.42	86.26	85.54	17.70	87.81	86.16	10.10	83.28	81.72	13.45	88.95	87.98	14.25
UNI-MMS	86.68	81.94	12.61	89.04	87.41	10.44	89.07	88.02	15.52	89.17	87.67	9.09	91.81	90.68	7.49	91.62	90.35	11.78
UNI-LB	87.96	82.98	10.28	84.15	82.57	14.33	90.21	89.72	13.16	90.38	89.17	7.20	86.88	85.68	11.71	93.21	91.56	10.28
UNI-IB	88.15	83.02	10.22	83.47	82.34	15.18	91.93	90.51	12.14	90.67	89.55	7.77	85.81	84.22	12.72	94.15	93.08	8.88
XLR-WSR	87.65	82.40	12.39	89.96	87.69	8.41	89.51	87.14	13.06	89.88	87.97	8.67	92.03	90.79	5.07	91.72	90.59	9.19
XLR-MMS	89.49	85.92	10.98	90.13	88.45	8.65	90.94	89.62	12.59	92.06	90.59	7.41	92.94	91.75	6.46	93.37	91.95	9.66
XLR-LB	91.62	88.21	6.65	**92.89**	**91.51**	7.05	92.89	90.92	11.44	94.35	92.68	2.68	94.99	**93.59**	4.37	95.75	94.50	9.06
XLR-IB	89.41	87.93	6.72	91.78	89.15	9.68	**94.14**	91.62	13.19	91.86	90.44	4.16	94.03	92.11	6.79	**97.00**	**95.45**	11.09
WSR-MMS	88.98	83.79	11.02	90.71	88.70	9.98	90.15	88.02	15.83	91.98	90.14	7.08	93.64	91.79	7.65	92.64	91.43	13.03
WSR-LB	89.59	84.18	8.20	91.72	89.52	7.24	90.12	89.04	12.05	91.65	90.27	4.23	93.99	92.21	4.33	92.54	90.58	**8.06**
WSR-IB	89.44	84.08	9.25	92.15	90.88	**6.92**	92.06	**91.72**	**11.21**	92.00	90.28	6.94	94.45	92.48	**3.76**	94.60	92.85	9.10
MMS-LB	**92.32**	86.94	**5.80**	92.15	90.22	8.59	92.29	91.51	11.41	**94.90**	**93.40**	**2.09**	94.31	92.35	6.43	94.48	92.88	**8.06**
MMS-IB	91.86	86.90	5.97	**93.20**	**91.15**	**6.70**	**94.23**	**92.18**	**10.69**	94.10	92.98	**2.19**	**95.45**	**93.86**	**3.46**	96.65	95.07	**7.64**
LB-IB	**93.40**	**90.15**	**5.48**	**94.22**	**92.98**	**6.42**	**95.50**	**93.28**	**8.30**	**96.33**	**95.13**	**2.05**	**96.46**	**95.06**	**3.08**	**97.80**	**96.48**	**4.80**

Table 4.5: Performance of PTM fusion using **MERLINN** compared with an E-CKA–only baseline.

Models trained on Previous SOTA NAC-generated speech detection datasets:We train the AASIST model—previously established as a SOTA benchmark for NAC-generated speech Wu *et al.* (2024*b*)—using one of the initial dataset introduced in Wu *et al.* (2024*b*). The trained model is then evaluated on our CFK dataset across C0, C1, and C2. The model attains accuracies of 67.08%, 65.33%, and 64.79%, along with corresponding EERs of 28.01%, 28.67%, and 29.45% on C0, C1, and C2, respectively. All models are trained using the same experimental configuration described earlier in the Training Pipeline section.

Comparison of PTMs with AASIST and CNN as Downstream Models: To validate our hypothesis, we evaluate a diverse set of PTMs using both AASIST and CNN as

downstream architectures. Models are trained on the respective training splits of the CFK dataset (C0, C1, and C2). The results, summarized in Table 4.4, *demonstrate that the multimodal PTMs—LB and IB—consistently achieve the strongest performance, thereby supporting our hypothesis*. These quantitative findings are further supported by t-SNE visualizations of the raw PTM representations shown in Figure 4.8, which reveal clearer clustering for multimodal PTMs. Notably, despite its simpler architecture, the CNN downstream model consistently outperforms AASIST across most PTM representations. Among monolingual speech PTMs, WavLM yields the best performance, whereas Wav2vec2 and Unispeech-SAT perform comparatively worse, highlighting their limited effectiveness in this setting. In contrast, multilingual speech PTMs—particularly XLS-R, Whisper, and MMS—consistently surpass monolingual counterparts, in line with observations from our SSD work Phukan *et al.* (2024*b*).

Results with `MERLINN`: The evaluation results of the models trained on fusion of PTM representations are shown in Table 4.5. As a baseline for comparison, we implement a euclidean-only CKA variant, where alignment is performed exclusively in euclidean space without incorporating the hyperbolic branch. To ensure fair evaluation, both the modeling architecture and training details are kept identical to those used for **`MERLINN`**. We compare the performance of PTM fusion against individual PTMs and observe that combining PTMs representations generally yields consistent improvements over individual PTM representatinos. The euclidean-only CKA baseline shows noticeable gains compared to individual PTMs, indicating the benefit of PTM fusion. However, **`MERLINN`** consistently delivers superior performance over both individual PTMs and the euclidean-only CKA baseline. By aligning representations across different geometric manifolds, **`MERLINN`** enables effective fusion of heterogeneous PTMs. In particular, fusing LB and IB using **`MERLINN`** achieves the best overall performance among all evaluated configurations. These results highlight the complementary nature of multimodal PTM representations and demonstrate **`MERLINN`**'s ability to effectively exploit this complementarity through its multi-geometry alignment strategy.

Comparison with SOTA: AASIST Wu *et al.* (2024*b*) represent SOTA architecture for NAC-generated speech detection. As discussed earlier, our results demonstrate that **`MERLINN`** consistently outperforms these approaches on the CCFD task, achieving the best overall performance.

Effectiveness of MERLINN on previous benchmark NAC-generated speech detection Datasets: To further assess the robustness of **MERLINN** alongside our best-performing configuration, we evaluate **MERLINN** with LB and IB on established benchmarks introduced by Wu *et al.* (2024*b*) and Lu *et al.* (2024). We train both AASIST and **MERLINN** on these datasets for a direct comparison. On the dataset from Lu *et al.* (2024), AASIST achieves an EER of 4.20%, whereas **MERLINN** reports a lower EER of 3.15%. Similarly, on the dataset proposed by Wu *et al.* (2024*b*), AASIST records an EER of 10.32%, while **MERLINN** improves upon this with an EER of 9.32%. These results demonstrate the superior performance and generalization capability of **MERLINN** compared to existing SOTA methods.

Statistical Significance: The statistical significance of the observed performance improvement was evaluated using a paired McNemar's test on the same test set. A pairwise comparison between **MERLINN** (LB + IB) and the strongest euclidean CKA baseline using the same PTMs resulted in p-values of 0.00854 (CO), 0.00473 (C1), and 0.00051 (C2). Since this value is well below the conventional significance threshold of 0.05, the performance gain achieved by **MERLINN** can be considered statistically significant.

4.4 Chapter Summary

This chapter investigated Geometry Guided Alignment strategies for fusion of heterogeneous PTM representations, demonstrating how geometric modeling can effectively reconcile representational disparities across diverse embedding spaces. We propose, three novel fusion framework, namely, **HYFuse** for SER, **MATA** for NVER, and **MERLINN** for SSD tackling the novel task of NAC-generated children speech detection. Through **HYFuse**, we showed that mapping heterogeneous representations into hyperbolic space preserves hierarchical acoustic–prosodic relationships. **MATA** further illustrated how optimal transport can align manifold structures across multimodal PTMs for improved NVER. Extending this geometric perspective, **MERLINN** unified euclidean and hyperbolic alignment to support SSD of vulnerable population i.e. Children. Collectively, these methods establish geometry as a principled and powerful foundation for achieving fusion across heterogeneous PTMs representation, forming the first major branch of the broader Proximal Interpolative Fusion paradigm for improved affective

and secure VIS.

CHAPTER 5

Interaction Guided Alignment

In this chapter, we introduce the Interaction Guided Alignment frameworks developed in this thesis. Prior studies have demonstrated that explicitly modeling interactions between heterogeneous PTM representations can substantially enhance downstream performance. For example, Arunkumar et al. Arunkumar *et al.* (2022) incorporated attention-based fusion to integrate complementary PTM representations for improved speech recognition. However, within the domains of affective and secure VIS, such interaction-driven fusion strategies remain largely unexplored. To address this gap, we propose two Interaction Guided Alignment frameworks. The first, **`MiO`**, leverages bilinear pooling to capture interactions among heterogeneous PTM representations, enabling richer cross-representation dependencies. The second, **`SCAR`**, employs a nested cross-attention mechanism to model bi-directional and hierarchical interactions between representations. We demonstrate that both approaches yield significant improvements for SSD, including challenging scenarios where the emotional content of the speech is intentionally manipulated.

5.1 MiO

In this section, we outline the background and motivation for leveraging interaction-based mechanism to fuse heterogeneous PTM representations, and subsequently introduce our proposed framework, **`MiO`** Phukan *et al.* (2024*a*), which utilizes bilinear pooling to model rich cross-representation interactions and achieve improved SSD performance.

5.1.1 Background and Motivation

With the growing availability of PTMs, SSD has seen rapid progress. Leveraging PTM-derived representations as input features offers significant benefits, including improved accuracy and reduced development effort compared to building SSD systems from

scratch. PTMs vary in architecture, pre-training strategies, and training data scale, and can be trained in supervised (e.g., Whisper Radford *et al.* (2023)) or self-supervised (e.g., Wav2vec2 Baevski *et al.* (2020)), as well as on single or multiple languages. Although typically pre-trained only on real speech, PTM representations have demonstrated strong performance in distinguishing genuine speech from synthetic counterparts wen Yang *et al.* (2021); Martín-Doñas and Álvarez (2022). However, previous research has overlooked the potential of multilingual speech PTMs for SSD. We believe that the representations from these PTMs will be the most effective for SSD. Our belief is grounded in the hypothesis that *multilingual PTMs, trained on large-scale and diverse multilingual speech data, inherently learn a broad range of pitch, accent, and tonal variations, making them more robust for SSD than other PTMs*. To test this, we evaluate eight SOTA PTMs—multilingual (XLS-R, Whisper, MMS), monolingual (Unispeech-SAT, WavLM, Wav2vec2), speaker recognition (x-vector), and emotion recognition (XLSR_emo) using two downstream networks: FCN and CNN across three benchmark datasets: ASV, ITW, and DECRO. Motivated by findings in other speech processing tasks Arunkumar *et al.* (2022) showing that certain PTM representations are complementary, we further propose a novel fusion framework, **M**erge **i**nto **O**ne (**MiO**), for combining PTM representations. In this section, we first identify the most effective PTM representations for SSD and subsequently design their fusion strategy. To our knowledge, this is the first study to explore PTM fusion for SSD. Our main contributions are:

- A comprehensive empirical analysis demonstrating the superior performance of multilingual PTMs for SSD, consistently outperforming other PTMs across all three datasets.
- Introducing **MiO**, a novel framework for fusion of heterogeneous PTM representations, which delivers significant gains over individual PTMs and achieves SOTA EER in ASV and ITW, with competitive results on DECRO.

5.1.2 Modeling

Modeling with Individual PTM representations: To evaluate the impact of PTM representations implicit knowledge learned for SSD, we keep the downstream models minimal. Two architectures are explored: (i) FCN applied directly to the extracted representations, and (ii) CNN: 1D-CNN followed by max pooling, an FCN block, and

a softmax output layer for probability estimation.

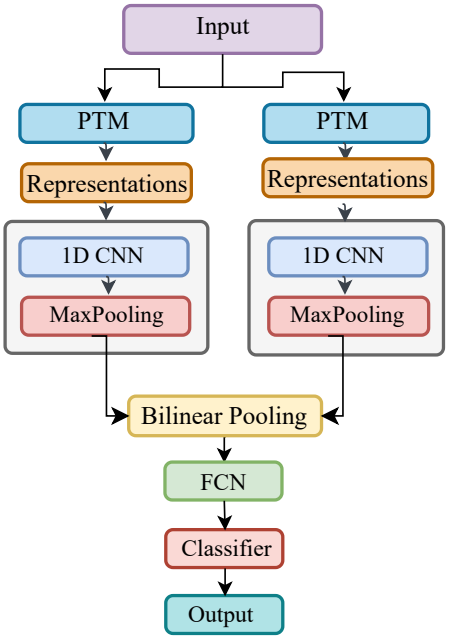


Figure 5.1: **MiO**

Merge into One (MiO): For fusion of representations from PTMs, we introduce **MiO** (Figure 5.1). Each input representation is first processed using the CNN-based approach from **Modeling with Individual PTM representations** (ii) above. Features are then linearly projected to a 120-dimensional space to reduce computational load, followed by bilinear pooling to capture multiplicative feature interactions. Bilinear pooling presents a effective and simple way to increase the interaction between the feature spaces Kumar and Nandakumar (2022). For vectors $\mathbf{u} \in \mathbb{R}^{D\times 1}$ and $\mathbf{v} \in \mathbb{R}^{D\times 1}$, bilinear pooling is defined as:

$$\mathbf{BP} = \mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^{\top} \tag{5.1}$$

The bilinear pooling output is flattened and passed through an FCN for final classification.

5.1.3 Experiments

Pre-Trained Models: We use XLS-R, Whisper, and MMS as multilingual PTMs which are SOTA in different multilingual speech processing tasks. As monolingual PTMs, we choose, WavLM (Large), WavLM (Base), Unispeech-SAT, and Wav2vec2. As speaker recognition and emotion recognition PTM, we choose, x-vector and XLSR_emo Cahyawijaya *et al.* (2023). These PTMs pre-trained for specialized objectives—such as speaker

recognition Ma *et al.* (2023) and SER Conti *et al.* (2022)—have demonstrated strong effectiveness for SSD, and as such we incorporate them in this thesis. XLSR_emo is an XLS-R-53 Conneau *et al.* (2021) model that has been further fine-tuned using the training splits of several English and Chinese SER corpora, including CREMA-D, CSED, ElderReact, ESD, IEMOCAP, and TESS. More details about the PTMs are given in Chapter 2.

Benchmark Datasets: For our experimental evaluation, we employed three widely recognized benchmark datasets: ASV, ITW, and DECRO. More details are given in Chapter 3.

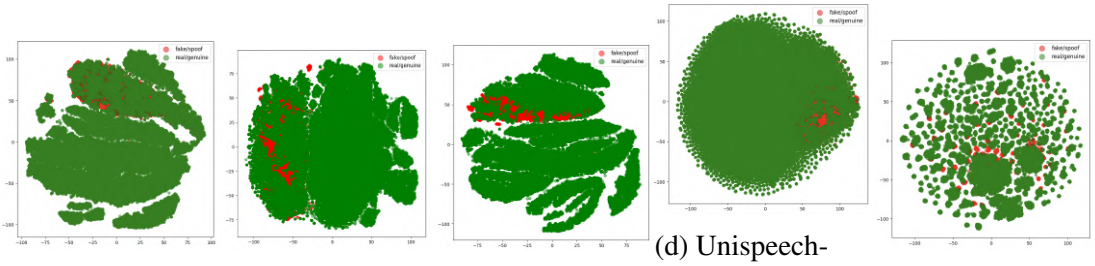
PTM	ASV	ITW	D-C	D-E
XLR	**1.67**	**0.24**	**1.42**	**0.12**
WSR	**2.34**	**0.73**	**0.69**	**0.11**
MMS	**3.10**	**0.31**	**1.11**	**0.19**
UNI	9.97	2.36	2.14	0.54
WLM (Base)	10.46	8.48	4.22	0.57
WLM (Large)	10.23	8.41	4.20	0.60
W2V	12.45	16.54	6.95	1.10
XVC	9.49	0.98	2.65	0.66
XLSR_emo	9.36	2.70	3.11	0.18

Table 5.1: EER (%) scores for FCN models with different PTM representations; D-C and D-E denote the Chinese and English subsets of DECRO, respectively. Abbreviations: W2V—Wav2vec2, WLM—WavLM, XVC—x-vector, XLR—XLS-R, UNI—Unispeech-SAT, WSR—Whisper

PTM	ASV	ITW	D-C	D-E
XLR	**1.03**	**0.17**	**1.42**	**0.07**
WSR	**2.02**	**0.22**	**0.58**	**0.06**
MMS	**1.50**	**0.20**	**0.70**	**0.09**
UNI	9.76	2.20	1.89	0.42
WLM (Base)	9.90	8.31	4.21	0.44
WLM (Large)	9.30	7.60	3.14	0.45
W2V	10.33	14.77	6.66	1.05
XVC	8.61	0.27	2.62	0.92
XLSR_emo	9.20	1.57	2.83	0.10

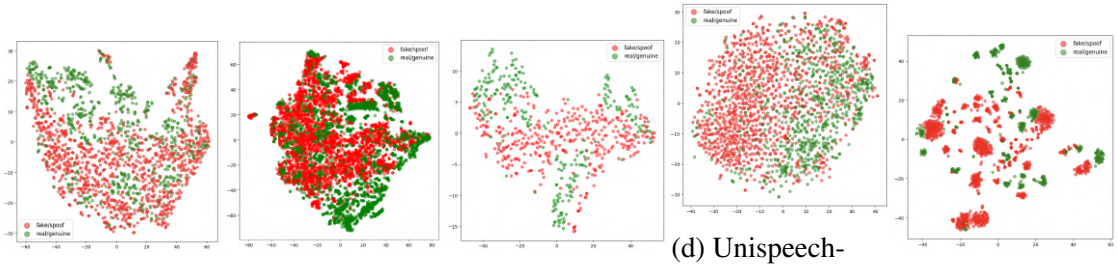
Table 5.2: EER (%) scores for CNN models with different PTM representations; D-C and D-E denote the Chinese and English subsets of DECRO, respectively.

Training Details: All the models are trained for 20 epochs with a batch size of 32, employing Adam optimization and cross-entropy loss. The learning rate is kept at 1e-3 and dropout regularization for reducing overfitting.



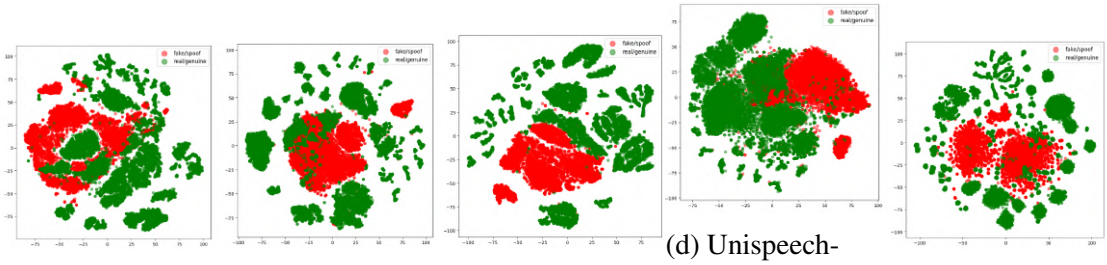
(a) XLS-R (b) Whisper (c) MMS (d) Unispeech-SAT (e) x-vector

Figure 5.2: t-SNE plots of raw representations from various PTMs on ASV



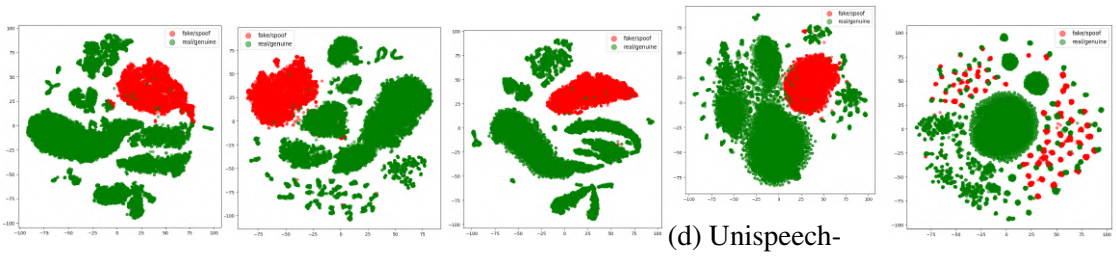
(a) XLS-R (b) Whisper (c) MMS (d) Unispeech-SAT (e) x-vector

Figure 5.3: t-SNE plots of raw representations from various PTMs on ITW



(a) XLS-R (b) Whisper (c) MMS (d) Unispeech-SAT (e) x-vector

Figure 5.4: t-SNE plots of raw representations from various PTMs on D-C



(a) XLS-R (b) Whisper (c) MMS (d) Unispeech-SAT (e) x-vector

Figure 5.5: t-SNE plots of raw representations from various PTMs on D-E

PTM Combinations	ASV	ITW	D-C	D-E
XLR + WSR	0.95	0.27	1.08	**0.05**
XLR + MMS	0.56	0.29	1.62	**0.06**
XLR + UNI	**0.45**	**0.11**	1.03	0.13
XLR + WLM (Base)	0.82	0.16	1.36	0.14
XLR + WLM (Large)	0.72	0.14	1.16	0.12
XLR + W2V	1.06	**0.12**	1.80	0.11
XLR + XVC	**0.41**	**0.07**	1.63	0.46
XLR + XLSR_emo	1.35	0.21	1.60	0.12
WSR + MMS	2.24	0.15	**0.27**	**0.08**
WSR + UNI	2.16	1.03	0.46	0.28
WSR + WLM (Base)	1.90	0.97	2.95	0.15
WSR + WLM (Large)	1.95	0.91	2.10	0.13
WSR + W2V	2.19	1.08	0.98	0.65
WSR + XVC	3.30	0.22	0.91	0.32
WSR + XLSR_emo	1.81	0.63	0.88	0.21
MMS + UNI	4.44	0.17	**0.19**	0.24
MMS + WLM (Base)	0.99	3.50	0.21	0.25
MMS + WLM (Large)	1.00	3.10	**0.15**	0.22
MMS + W2V	**0.50**	0.22	0.39	0.33
MMS + XVC	5.40	0.14	0.77	0.25
MMS + XLSR_emo	1.80	0.36	0.81	0.32
UNI + WLM (Base)	10.18	2.79	2.31	0.48
UNI + WLM (Large)	9.19	2.99	2.11	0.41
UNI + W2V	9.74	2.55	2.88	0.59
UNI + XVC	5.82	0.15	2.56	0.54
UNI + XLSR_emo	6.80	1.70	2.09	0.57
WLM (Base) + W2V	12.46	8.54	1.93	0.51
WLM (Base) + XVC	6.03	0.19	3.04	0.61
WLM (Base) + XLSR_emo	7.91	2.31	2.64	0.21
WLM (Large) + W2V	11.36	7.14	1.44	0.54
WLM (Large) + XVC	5.01	0.19	2.21	0.60
WLM (Large) + XLSR_emo	6.92	2.20	2.01	0.18
W2V + XVC	7.31	0.26	3.51	0.47
W2V + XLSR_emo	7.50	1.91	2.87	0.32
XVC + XLSR_emo	7.89	0.43	1.11	0.71

Table 5.3: EER (%) scores for different PTM representation combinations with **MiO**. Abbreviations: W2V—Wav2vec2, WLM—WavLM, XVC—x-vector, XLR—XLS-R, UNI—Unispeech-SAT, WSR—Whisper. D-C and D-E denote the Chinese and English subsets of DECRO, respectively.

Experimental Results: For ASV and DECRO, results are reported on the official test split, while ITW scores are averaged over five random splits due to the absence of an official partition. DECRO models are trained separately on Chinese (D-C) and English (D-E) subsets. We have used EER as the evaluation metric as used by previous research Martín-Doñas and Álvarez (2022). Tables 5.1 and 5.2 show EERs for dif-

ferent PTM representations using FCN and CNN downstream models. Multilingual PTMs (XLS-R, Whisper, MMS) consistently achieve the lowest EERs, *validating our hypothesis that exposure to diverse multilingual data equips them with richer pitch, accent, and tone variations beneficial for SSD*. XLS-R attains the best scores for ASV and ITW, while Whisper leads in DECRO. MMS exhibits mixed but competitive performance. CNN models generally outperform FCNs by capturing richer local patterns. Among monolingual PTMs, Unispeech-SAT stands out, outperforming WavLM (Large) despite having the same parameter count as WavLM (Base), likely due to its speaker-aware pre-training. Specialized models such as x-vector and XLSR_emo also surpass general monolingual PTMs in some cases, owing to their task-specific feature learning. Wav2vec2 performs worst, indicating its limited ability to capture features critical for fake–real discrimination. We have also presentations visualizations of the PTM representation spaces in Figures 5.2, 5.3, 5.4, and 5.5 for ASV, ITW, D-C, and D-E, respectively. They show that multilingual PTMs exhibit clearer class separation between real and fake samples and thus augmenting our results. Table 5.3 presents results from PTM fusion. Combining XLS-R with x-vector yields the lowest EERs for ASV and ITW, suggesting strong complementarity between multilingual and speaker-specific features. In D-C, MMS with WavLM (Large) performs best, while D-E benefits from XLS-R with Whisper. However, some combinations degrade performance, e.g., XLS-R with XLSR_emo, where contradictory feature behaviors emerge.

Dataset	Model	EER (%)
ASV	CQT-DCT-LCNN Lavrentyeva *et al.* (2019)	**1.84**
	MiO(XLR + XVC)	**0.41**
ITW	STATNet Ranjan *et al.* (2022)	**0.20**
	MiO(XLR + XVC)	**0.07**
D-E	Res-TSSDNet Ba *et al.* (2023)	**0.02**
	MiO(XLR + WSR)	**0.04**

Table 5.4: Comparison with SOTA works on ASV, ITW, and D-E in terms of EER (%). **MiO(XLR + XVC)** and **MiO(XLR + WSR)** denote the proposed **MiO** framework instantiated with the corresponding PTM pair, where XLR, XVC, and WSR refer to XLS-R, x-vector, and Whisper, respectively.

Comparison with State-of-the-art: We benchmark **`MiO`** against prior SOTA methods on ASV, ITW, and D-E, as shown in Table 5.4. Since Ba *et al.* (2023) used D-C solely for cross-lingual evaluation from English to Chinese, no prior works report training and

PTM	ASV Training			D-C Training			ITW Training			D-E Training		
	D-C	D-E	ITW	ASV	ITW	D-E	ASV	D-C	D-E	ASV	D-C	ITW
XLR	**22.78**	**12.21**	**28.11**	**10.23**	**19.78**	**15.32**	**10.45**	**18.88**	**21.03**	**11.06**	**39.58**	**14.94**
WSR	**30.51**	**11.09**	**29.21**	**17.19**	**35.55**	**15.34**	**16.91**	**39.62**	**30.30**	**15.78**	**30.67**	**17.18**
MMS	**19.62**	**9.61**	**18.83**	**4.11**	**19.51**	**8.71**	27.50	50.63	**20.19**	**14.78**	**32.75**	**21.59**
W2V	48.23	19.48	43.01	19.82	40.07	18.74	45.37	49.15	49.99	28.80	46.21	35.03
WLM (Large)	31.10	13.19	32.12	31.23	36.69	15.99	28.13	**43.91**	35.93	16.61	40.60	24.61
XVC	32.98	13.62	37.43	39.53	38.66	17.69	**25.43**	47.90	37.95	18.62	39.65	28.62

Table 5.5: Cross-corpus evaluation using 120-dimensional PTM representations; results are reported in EER (%). Training is performed on one corpus (ASV, D-C, ITW, or D-E), and evaluation is conducted on the remaining corpora (shown as columns). For ITW, a single evaluation split is used based on one random seed. Abbreviations: XLR—XLS-R, WSR—Whisper, W2V—Wav2vec2, WLM-WavLM, and XVC—x-vector.

testing directly on D-C. Our approach achieves the lowest EER on ASV and ITW, and delivers competitive results on D-E compared to existing best-performing methods.

PTM	ASV Training			D-C Training			ITW Training			D-E Training		
	D-C	D-E	ITW	ASV	ITW	D-E	ASV	D-C	D-E	ASV	D-C	ITW
XLR	**21.33**	**12.78**	**29.28**	**12.14**	**23.22**	**21.88**	**8.21**	**17.69**	**29.88**	**21.99**	**11.93**	**14.65**
WSR	**25.22**	**21.00**	**34.88**	**13.66**	**29.22**	**16.94**	**17.82**	**33.60**	**17.87**	**24.88**	**11.87**	**21.86**
MMS	**34.61**	**26.14**	**19.28**	**19.88**	**31.40**	**32.97**	**20.74**	42.11	**10.85**	**23.16**	**18.42**	**21.43**
W2V	57.66	43.98	46.66	44.63	47.22	41.90	27.34	53.33	44.97	35.76	43.98	46.95
WLM (Large)	35.19	30.31	42.20	34.99	33.21	34.28	21.11	**41.39**	30.83	29.51	33.47	35.11
XVC	35.79	32.34	45.04	37.09	39.22	39.02	21.12	43.22	31.00	32.55	32.77	32.98

Table 5.6: Cross-corpus evaluation using 240-dimensional PTM representations; results are reported in EER (%). Training is performed on one corpus (ASV, D-C, ITW, or D-E), and evaluation is conducted on the remaining corpora (shown as columns). For ITW, a single evaluation split is used based on one random seed. Abbreviations: XLR—XLS-R, WSR—Whisper, W2V—Wav2vec2, WLM-WavLM, and XVC—x-vector.

Model	ASV Training			D-C Training			ITW Training			D-E Training		
	D-C	D-E	ITW	ASV	ITW	D-E	ASV	D-C	D-E	ASV	D-C	ITW
XLR + XVC	**21.11**	**12.04**	**24.80**	**16.53**	**23.34**	**16.14**	**58.78**	**49.84**	**69.14**	**21.34**	**35.72**	**48.15**
WSR + UNI	**44.59**	**7.80**	**38.86**	**26.89**	**47.02**	**15.74**	**55.02**	**29.94**	**47.72**	**12.89**	**44.26**	**27.50**

Table 5.7: Cross-corpus evaluation using fused PTM representations reduced to 120 dimensions; results are reported in EER (%). Training is performed on one corpus (ASV, D-C, ITW, or D-E), and evaluation is conducted on the remaining corpora (shown as columns). For ITW, the test set corresponds to a single split generated from one random seed. Abbreviations: XLR—XLS-R, WSR—Whisper, W2V—Wav2vec2, and XVC—x-vector.

Model	ASV Training			D-C Training			ITW Training			D-E Training		
	D-C	D-E	ITW	ASV	ITW	D-E	ASV	D-C	D-E	ASV	D-C	ITW
XLR + XVC	**17.21**	**15.14**	**24.53**	**14.84**	**17.97**	**13.53**	**55.85**	**50.91**	**57.04**	**37.68**	**14.23**	**26.91**
WSR + UNI	**31.21**	**15.70**	**41.56**	**24.61**	**46.94**	**18.52**	**54.80**	**39.57**	**46.75**	**42.57**	**18.34**	**31.09**

Table 5.8: Cross-corpus evaluation using fused PTM representations reduced to 240 dimensions; results are reported in EER (%). Training is performed on one corpus (ASV, D-C, ITW, or D-E), and evaluation is conducted on the remaining corpora (shown as columns). For ITW, the test set corresponds to a single split generated from one random seed.

Statistical Significance: To evaluate the statistical reliability of the reported performance gains, paired McNemar's tests were performed using predictions obtained from identical test sets across ASV, ITW, DECRO-Chinese and DECRO-English. The proposed **MiO** framework, combining Whisper and MMS representations, was compared against the strongest competing baseline based on feature concatenation of the same PTMs. The resulting p-values were 0.00017 for ASV, 0.00652 for ITW, 0.00547 for DECRO-Chinese and 0.00310 for DECRO-English, all of which are below the conventional significance threshold of 0.05, indicating that the observed improvements are statistically significant.

Cross-Corpus Evaluation: We also evaluate how well dowstream models trained on multilingual PTM representations transfer to unseen corpora—a setting where prior SSD studies Ba *et al.* (2023); Müller *et al.* (2022*a*) have shown that systems trained on one dataset or language often struggle when tested on another. Using the same downstream architecture, we first normalize representation dimensionality across PTMs using Principal Component Analysis (PCA), projecting all representations to either 120 or 240 dimensions. This projection step is necessary because the original PTM representations differ in dimensionality, and it allows us to assess their intrinsic behavior under cross-corpus evaluation conditions. Models are then trained on the training split of one dataset and evaluated on the test split of each remaining dataset. Training hyperparameters (epochs, batch size, and optimization setup) follow the configuration described in Training Details above. For a meaningful comparison, we benchmark multilingual PTMs against a monolingual PTM (Wav2vec2) as well as a strong speaker-recognition PTM (x-vector), both of which produced competitive performance in earlier experiments (Table 5.2). Cross-corpus findings (Tables 5.5 and 5.6) consistently show that multilingual PTMs representations offer superior transferability across datasets. Although multilingual PTMs generally lead, their performance varies across settings—for example, MMS achieves the strongest results when trained on ASV and D-C in the 120-dimensional configuration, whereas XLS-R performs competitively when trained on ITW and D-E and evaluated on the remaining corpora. We also observe noticeable shifts between the 120- and 240-dimensional PCA projections, indicating that representational dimensionality has a subtle but measurable influence on downstream generalization. Complementary cross-corpus evaluations for selected PTM pairs using **MiO** (Tables 5.7 and 5.8) further demonstrate that fusing top-performing PTM representa-

tions through **MiO** enhances robustness across datasets, surpassing the generalization achieved by their individual PTM representations.

5.2 SCAR

In this section, we discuss the background and motivation for leveraging multimodal PTMs representations for an edge case of synthetic speech—specifically when the emotional content of the speech is deliberately altered. We then introduce our proposed framework, **SCAR**, situated within the broader family of Interaction Guided Alignment methods. Unlike **MiO**, which models cross-representation relationships through bilinear pooling, **SCAR** Akhtar *et al.* (2025) employs a nested cross-attention mechanism to capture fine-grained, bi-directional dependencies between heterogeneous PTM representations. By recursively modeling how one PTM attends to another in a nested fashion interaction depths, **SCAR** learns rich relationship that are essential for identifying emotion-manipulated synthetic speech.

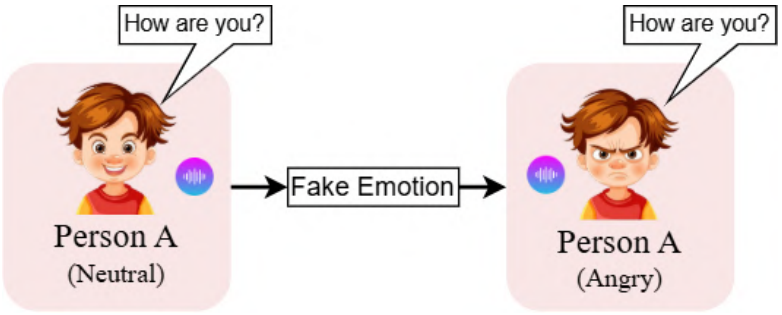


Figure 5.6: Illustration of EmoFake: Speaker A's happy speech is altered to synthesize a sad emotional tone while preserving the original spoken content

5.2.1 Background and Motivation

With the increasing accessibility of powerful PTMs, research in SSD has progressed considerably Guo *et al.* (2024); Pascu *et al.* (2024); Yang *et al.* (2024). PTMs trained on large, diverse speech corpora have driven much of this progress, offering substantial performance gains while eliminating the need for task-specific training from scratch—a trend observed in both general SSD Kawa *et al.* (2023) and also in synthetic singing voice detection Chen *et al.* (2024). However, despite these advancements, one critical vulnerability remains largely overlooked: synthetic speech in which the emotional

content is intentionally altered. For convenience, we refer to this class of attacks as EmoFake throughout this section.

EmoFake (Figure 5.6) involves modifying a speaker's emotional profile—such as pitch, timbre, and intensity—while keeping both the linguistic content and speaker identity intact. Unlike conventional synthetic speech (Detecting conventional synthetic speech is handled by **MiO** in the above Section), EmoFake performs targeted emotional manipulation, making the altered speech perceptually natural yet deceptive. This subtle form of attack introduces new security and forensic challenges. They can be exploited to induce affective bias in affect-aware VIS, increasing susceptibility to persuasion, coercion, or inappropriate system actions. As a remedy, Zhao et al. Zhao *et al.* (2024*b*) introduced the first dataset for detecting such synthetic speech and provided initial baseline results. However, their work did not explore the use of large-scale PTMs, even though such models have consistently shown strong effectiveness across a wide range of SSD tasks. In this chapter, we focus specifically on EmoFake Detection (EFD)[1] and undertake, to the best of our knowledge, the first comprehensive evaluation of PTMs for EFD. *We hypothesize that multimodal PTMs—(LB and IB)—will surpass speech-only PTMs (WavLM, Wav2vec2, UniSpeech-SAT) in EFD. Unlike speech PTMs, multimodal PTMs undergo cross-modal pre-training and thus learn richer emotional structures shared across modalities.* This cross-modal grounding allows them to better recognize unnatural emotional shifts and inconsistencies, making them more capable of discriminating authentic from manipulated emotional expressions. To evaluate this hypothesis, we perform an extensive comparison of SOTA speech and multimodal PTMs. Our findings show that multimodal PTMs consistently outperform speech PTMs on EFD.

Further, motivated by our prior work in SSD Phukan *et al.* (2024*a*) (**MiO**) and SER Wu *et al.* (2023*a*), where fusing heterogeneous PTMs representations yielded notable gains, we extend this line of inquiry to EFD. To the best of our knowledge, this is the first study to explore fusion of PTM representations for EFD. To this end, we introduce **SCAR** (Ne**S**ted **C**ross-**A**ttention Netwo**R**k), a novel Interaction Guided Alignment framework designed for effective fusion of PTM representations. **SCAR** incorporates a nested cross-attention mechanism that enables multi-stage, bidirectional information

[1]* To clearly distinguish this setting from conventional SSD, we refer to this task as EFD throughout this section

exchange between PTMs, allowing the model to progressively refine emotional cues across representational spaces. A complementary self-attention refinement module further amplifies cross-PTM patterns while suppressing irrelevant variations. By applying **SCAR** to fuse multimodal PTMs, we obtain the strongest performance—surpassing standalone PTMs, baseline fusion technique, and previous SOTA systems—thereby establishing a new SOTA for EFD. Similar to the fusion approaches introduced earlier in Chapter 4—namely **MATA** and Chapter 5 **MiO**—this section follows the same methodology: we first identify the most effective PTMs for the downstream task and then formulate their corresponding fusion strategy to achieve improved performance.

The key contributions are given as follows:

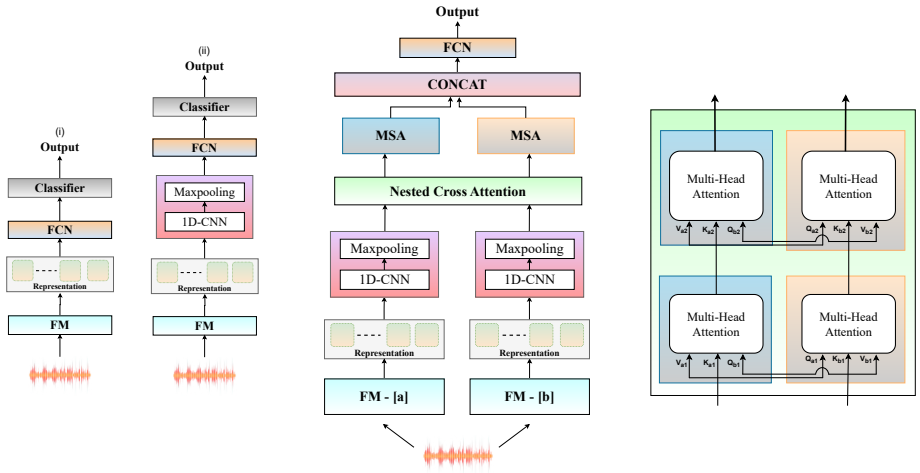
- We conduct the first extensive comparative study of SOTA multimodal and speech PTMs to assess their suitability for EFD. Our findings demonstrate that multimodal PTMs consistently deliver superior performance.
- We introduce **SCAR**, a novel framework for fusion of heterogeneous PTMs. By applying **SCAR** to combine multimodal PTMs, we achieve performance that surpasses individual PTMs, outperforms baseline fusion approach, and establishes new SOTA results compared to prior work.

5.2.2 Modeling

We first provide a detailed description of the downstream modeling used with individual PTMs, followed by our proposed fusion framework, **SCAR**.

Modeling with Individual PTM representations: We adopt FCN and CNN as downstream classifiers, consistent with prior work in related domains such as SER Chetia Phukan *et al.* (2023) and shout-intensity estimation Fukumori *et al.* (2023). The CNN architecture (Figure 5.7 (b)) begins with a 1D convolutional layer containing 32 filters of kernel size 3, followed by a max-pooling operation. The resulting feature maps are flattened and passed through a two-layer FCN comprising 512 and 128 units, respectively, before reaching the final sigmoid-activated output layer. The FCN baseline mirrors the fully connected block used in the CNN, without the convolutional front-end. Depending on the dimensionality of the PTM representation, the CNN models contain approximately 0.5–0.7M trainable parameters, whereas the FCN models range between 0.45–0.6M parameters.

SCAR: Framework: The architecture of the proposed framework is illustrated in



(a) (b) (c)

Figure 5.7: Modeling architectures: (a) FCN and CNN; (b) **SCAR**; and (c) a detailed illustration of the nested cross-attention mechanism

Figure 5.7. **SCAR** introduces a nested cross-attention mechanism that enables multi-stage interactions between the representations extracted from PTMs, allowing progressively refined information exchange. Cross-attention has proven to be an effective fusion mechanism across a wide range of multimodal tasks for integrating different modalities Ilias and Askounis (2024); Praveen and Alam (2024). However, its application within affective and secure VIS—particularly in a nested cross-attention formulation—has not been explored previously. Further, we add a subsequent self-attention refinement module further enhances discriminative cues by amplifying cross-PTM patterns while suppressing noise. First, representations obtained from the selected PTMs are passed through a convolutional front-end identical to that used in individual PTM modeling. The flattened outputs are then fed into the nested cross-attention block, which consists of two sequential cross-attention stages. Let the flattened representations from two PTMs P and Q be denoted by $\mathbf{X}_P$ and $\mathbf{X}_Q$, respectively. For the first cross-attention stage, the query ($\mathbf{A}$), key ($\mathbf{B}$), and value ($\mathbf{C}$) projections are computed as:

$$\mathbf{A}_P^{(1)} = \mathbf{X}_P \mathbf{U}_{A1}^P, \qquad \mathbf{B}_Q^{(1)} = \mathbf{X}_Q \mathbf{U}_{B1}^Q, \qquad \mathbf{C}_Q^{(1)} = \mathbf{X}_Q \mathbf{U}_{C1}^Q, \tag{5.2}$$

$$\mathbf{A}_Q^{(1)} = \mathbf{X}_Q \mathbf{U}_{A1}^Q, \qquad \mathbf{B}_P^{(1)} = \mathbf{X}_P \mathbf{U}_{B1}^P, \qquad \mathbf{C}_P^{(1)} = \mathbf{X}_P \mathbf{U}_{C1}^P. \tag{5.3}$$

The corresponding cross-attention outputs are obtained as:

$$\mathbf{H}_P^{(1)} = \text{softmax}\left(\frac{\mathbf{A}_P^{(1)}\mathbf{B}_Q^{(1)\top}}{\sqrt{d_a}}\right)\mathbf{C}_Q^{(1)}, \tag{5.4}$$

$$\mathbf{H}_Q^{(1)} = \text{softmax}\left(\frac{\mathbf{A}_Q^{(1)}\mathbf{B}_P^{(1)\top}}{\sqrt{d_a}}\right)\mathbf{C}_P^{(1)}. \tag{5.5}$$

For the second cross-attention stage, updated projections are computed as:

$$\mathbf{A}_P^{(2)} = \mathbf{H}_P^{(1)}\mathbf{U}_{A2}^P, \qquad \mathbf{B}_Q^{(2)} = \mathbf{H}_Q^{(1)}\mathbf{U}_{B2}^Q, \qquad \mathbf{C}_Q^{(2)} = \mathbf{H}_Q^{(1)}\mathbf{U}_{C2}^Q, \tag{5.6}$$

$$\mathbf{A}_Q^{(2)} = \mathbf{H}_Q^{(1)}\mathbf{U}_{A2}^Q, \qquad \mathbf{B}_P^{(2)} = \mathbf{H}_P^{(1)}\mathbf{U}_{B2}^P, \qquad \mathbf{C}_P^{(2)} = \mathbf{H}_P^{(1)}\mathbf{U}_{C2}^P. \tag{5.7}$$

The refined cross-attention outputs are then given by:

$$\mathbf{H}_P^{(2)} = \text{softmax}\left(\frac{\mathbf{A}_P^{(2)}\mathbf{B}_Q^{(2)\top}}{\sqrt{d_a}}\right)\mathbf{C}_Q^{(2)}, \tag{5.8}$$

$$\mathbf{H}_Q^{(2)} = \text{softmax}\left(\frac{\mathbf{A}_Q^{(2)}\mathbf{B}_P^{(2)\top}}{\sqrt{d_a}}\right)\mathbf{C}_P^{(2)}. \tag{5.9}$$

Finally, a self-attention refinement step is applied to each PTM stream:

$$\mathbf{R}_P = \text{softmax}\left(\frac{\mathbf{H}_P^{(2)}\mathbf{H}_P^{(2)\top}}{\sqrt{d_a}}\right)\mathbf{H}_P^{(2)}, \tag{5.10}$$

$$\mathbf{R}_Q = \text{softmax}\left(\frac{\mathbf{H}_Q^{(2)}\mathbf{H}_Q^{(2)\top}}{\sqrt{d_a}}\right)\mathbf{H}_Q^{(2)}. \tag{5.11}$$

The final fused representation is obtained through concatenation:

$$\mathbf{X}_{\text{fused}} = \text{Concat}(\mathbf{R}_P, \mathbf{R}_Q). \tag{5.12}$$

The fused output is then passed through a FCN comprising two dense layers of 512 and 128 units, respectively, followed by output layer with sigmoid activation for binary classification. Both cross-attention stages use two attention heads, and the self-attention refinement also employs two heads. Depending on the dimensionality of the PTM representations, **SCAR** contains approximately 1.8-2.3M trainable parameters.

5.2.3 Experiments

Pre-Trained Models: We use Unispeech-SAT, Wav2vec2, WavLM (Base), HuBERT, and Whisper as speech PTMs. As multimodal PTMs, we use LB and IB. Details about the PTMs can be in Chapter 2.

Dataset: We use the official Emofake split for training, validation, and model evaluation. More detailed information can be found in Chapter 3.

Training Details: We employ Adam optimizer to train the models. We keep learning rate of 1e-3 and a batch size of 32. We use binary cross-entropy as loss function and to prevent overfitting, we implement dropout.

	E		C	
PTM	**FCN**	**CNN**	**FCN**	**CNN**
UNI	2.43	2.21	7.68	7.35
W2V	2.51	2.24	4.36	3.29
WLM	5.67	5.25	8.78	4.91
HBT	3.62	2.93	10.94	10.23
WSR	2.31	2.04	3.25	3.11
LB	**1.64**	**1.32**	**2.19**	**1.89**
IB	**1.87**	**1.36**	**2.10**	**1.81**

Table 5.9: EER (%) for each PTM on the English (E) and Chinese (C) subsets using FCN and CNN back-ends. Abbreviations: W2V—Wav2vec2, WLM—WavLM, UNI—Unispeech-SAT, HBT—HuBERT, WSR—Whisper. Darker color tones denote lower EER (better), while lighter tones indicate higher EER. The same PTM abbreviations and color-coding scheme are used in Table 5.10.

Experimental Results: We use EER as the evaluation metric, following previous research on EFD Zhao *et al.* (2024*b*). Table 5.9 displays the evaluation scores of downstream models trained on representations from different PTMs. Our findings show that multimodal PTMs consistently achieve lower EER values across both language subsets (English and Chinese) and downstream network architectures when compared to unimodal speech PTMs. *This supports our hypothesis that multimodal PTMs will outperform speech PTMs in EFD, as their cross-modal pretraining enables them to capture emotional patterns from multiple modalities, improving their ability to detect unnatural emotional shifts and inconsistencies in manipulated speech.* Among speech PTMs, Whisper achieved the best performance, likely due to its multilingual pretraining. This is consistent with prior research on SSD Phukan *et al.* (2024*a*) (Chapter 5 **MiO**), which

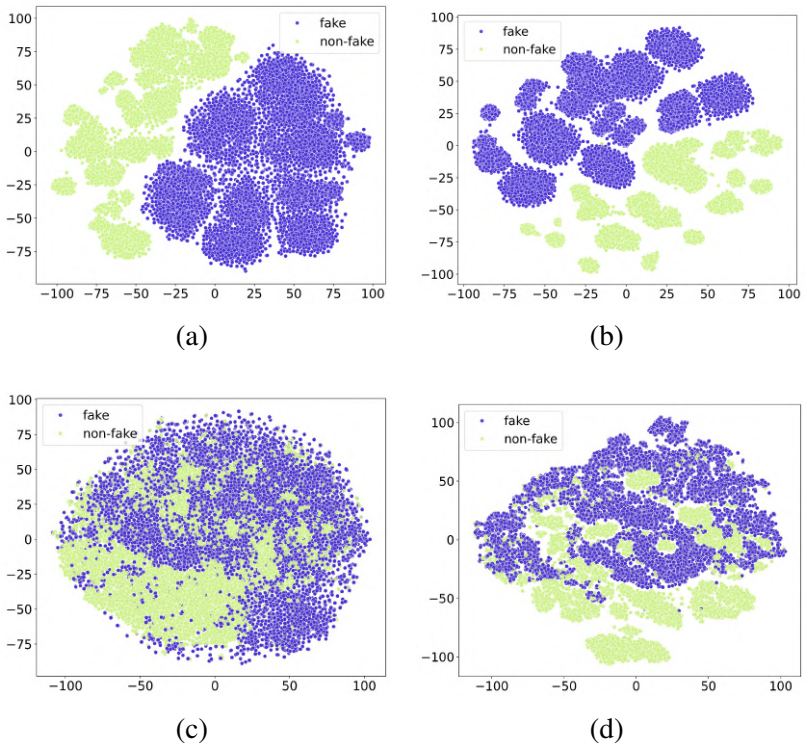


Figure 5.8: Subfigures (a) LB, (b) IB, (c) Wav2vec2, and (d) Whisper show the corresponding t-SNE visualizations

showed that multilingual speech PTMs are better at capturing variations in pitch, tone, and intensity, thus enhancing their ability to detect synthetic speech. Overall, CNN-based models generally achieve lower EER scores than FCNs across all PTMs. We also include t-SNE plots of raw representations from the PTMs in Figure 5.8, where clear class clusters are visible for multimodal PTMs, further supporting our findings.

Table 5.10 presents the evaluation scores for various combinations of PTMs. We use concatenation as the baseline fusion technique. For the baseline, we adopt a similar architecture to **SCAR**, omitting the nested cross-attention block and self-attention refinement. The training specifications follow those used for **SCAR**. We observe that fusion of PTMs through **SCAR** consistently outperforms the concatenation based approach, highlighting the effectiveness of **SCAR** for fusion. The best performance is achieved by fusing multimodal PTMs, specifically LB and IB, using **SCAR**, demonstrating the strong complementary relationship between these modalities. This underscores the significance of cross-interaction among features in a nested fashion for enhancing performance in EFD, further validating the effectiveness of interaction guided alignment of PTMs.

Comparison to SOTA: We compare our best model, **SCAR** with LB and IB, to the

Combinations	Concatenation		**SCAR**	
	E	**C**	**E**	**C**
UNI + W2V	3.93	6.63	2.45	4.05
UNI + WLM	3.48	7.05	3.61	6.56
UNI + HBT	3.06	5.21	2.96	4.62
UNI + WSR	4.11	4.35	2.12	2.84
UNI + LB	3.42	2.94	1.95	2.45
UNI + IB	2.90	3.43	1.89	1.97
W2V + WLM	2.65	5.14	2.13	3.94
W2V + HBT	3.04	5.87	2.96	5.31
W2V + WSR	3.03	6.11	2.63	3.07
W2V + LB	2.07	2.93	1.92	1.10
W2V + IB	2.35	4.94	1.85	2.04
WLM + HBT	3.01	4.34	2.86	3.57
WLM + WSR	3.76	4.06	3.63	2.92
WLM + LB	2.98	3.92	1.95	2.94
WLM + IB	2.93	3.37	2.13	2.45
HBT + WSR	3.21	4.86	2.65	3.02
HBT + LB	4.04	6.02	3.02	2.93
HBT + IB	3.03	5.07	2.73	4.90
WSR + LB	1.38	1.89	1.24	1.66
WSR + IB	2.67	1.93	1.53	1.81
LB + IB	**1.21**	**1.43**	**1.15**	**1.02**

Table 5.10: Evaluation results (EER in %) for pairwise PTM combinations using concatenation and **SCAR**; E and C denote the English and Chinese subsets, respectively.

previous SOTA work Zhao *et al.* (2024*b*), which reported EER values of 3.65% and 8.34% for the English and Chinese subsets, respectively. In contrast, **SCAR** with LB and IB achieved the lowest EER values of 1.15% and 1.02% for English and Chinese, respectively, setting a new SOTA for EFD.

Statistical Significance: The statistical significance of the observed performance improvement was evaluated using a paired McNemar's test on the same test set. A comparison between **SCAR** (LB + IB) and the strongest concatenation-based (LB + IB) baseline using the same PTMs yielded p-values of 0.00185 (English) and 0.00037 (Chinese), which is below the conventional significance threshold of 0.05, indicating that the performance gain is statistically significant.

5.3 Chapter Summary

In this chapter, we have proposed introduced the Interaction Guided Alignment frameworks developed in this thesis, aiming to bridge a key gap in the fusion of heterogeneous PTM representations for affective and secure VIS. While prior work has shown the benefits of modeling interactions between different PTM representations, such strategies have been limited in the domains we address. To fill this gap, we proposed two novel frameworks: **`MiO`**, which uses bilinear pooling to capture inter-PTM interactions, and **`SCAR`**, which employs a nested cross-attention mechanism to model interactions. Both frameworks demonstrated substantial improvements in SSD including in challenging cases where emotional content in speech is manipulated (Emofake). These results highlight the effectiveness of Interaction Guided Alignment in enhancing performance and setting the stage for future advancements in the field.

CHAPTER 6

Divergence Guided Alignment

In this chapter, we introduce the Divergence Guided Alignment frameworks developed in this thesis. Prior research in multimodal learning has shown that aligning representations at the distributional level—rather than only at the feature or interaction level—can significantly improve downstream performance Shankar (2022); Wan *et al.*. However, such fusion approaches have been rarely used in affective and secure VIS applications. To bridge this gap, we propose two Divergence Guided Alignment frameworks. The first, **FINDER**, employs renyi divergence to align the distributional characteristics of heterogeneous PTM representations, enabling effective SSSA. The second, **COFFE**, leverages chernoff distance to optimize distributional alignment, particularly targeting challenging scenarios such as SSSA of synthetic singing voice. Through extensive experiments, we demonstrate that Divergence Guided Alignment substantially enhances performance for SSSA under diverse conditions.

6.1 FINDER

In this section, we present the background and motivation followed by highlighting the role of prosodic characteristics as key fingerprints for effective SSSA. We then introduce **FINDER** Phukan *et al.* (2025*f*), a novel Divergence Guided Alignment framework designed to enhance SSSA performance.

6.1.1 Background and Motivation

Prior research in the synthetic speech mitigation community has predominantly concentrated on binary classification—distinguishing genuine from synthetic speech (SSD) Wu *et al.* (2015*b*); Kinnunen *et al.* (2017); Todisco *et al.* (2019); Liu *et al.* (2023); Yamagishi *et al.* (2021); Yi *et al.* (2022, 2023); Shaaban *et al.* (2023); Hamza *et al.* (2022); Altalahin *et al.* (2023); Kilinc and Kaledibi (2023). Although this formulation

is effective due to its simplicity, it falls short in addressing a critical dimension of synthetic speech analysis: source attribution. SSSA extends beyond the binary real–fake paradigm by aiming to identify the specific generative tool or model responsible for producing synthetic speech Yan *et al.* (2022*b*); Zhang *et al.* (2023); Zhu *et al.* (2022). Such fine-grained attribution plays a vital role by improving the interpretability of detection decisions and facilitating targeted countermeasures, especially in sensitive applications including audio forensics, legal proceedings, and intellectual property enforcement. Modern generative frameworks, including TTS and VC systems, inherently imprint distinctive prosodic characteristics—such as pitch dynamics, timbre, rhythm, and intonation—onto the synthesized speech. These attributes arise from the architectural design choices and training pipelines of the generative models and thus act as prosodic fingerprints that are instrumental for effective SSSA. However, no previous research in SSSA, has focused on such fingerprints. In this work, for the first time, we systematically investigate a range of SOTA speech PTMs to assess their ability to capture such source-specific prosodic signatures for SSSA. We focus on PTMs that have demonstrated strong performance in prosody-sensitive tasks, including SER and depression detection. Our evaluation is motivated by prior studies showing the effectiveness of PTMs in both SSSA-related and SSD tasks Klein *et al.* (2024); Phukan *et al.* (2024*a*), where they provide substantial performance gains while eliminating the need for training models from scratch. Furthermore, inspired by our works on SSD Phukan *et al.* (2024*a*) and NVER Phukan *et al.* (2025*e*), we explore the fusion of heterogeneous PTM representations and introduce **FINDER** (**F**us**I**on through Re**N**yi **D**iv**ER**gence), a Divergence Guided Alignment framework designed for effective representation alignment for improved SSSA. To the best of our knowledge, this work constitutes the first systematic investigation of PTM representations fusion for SSSA. Beyond its technical contributions, our framework directly benefits end users by enabling more transparent and explainable affective and secure VIS, empowering users to better understand not only whether speech is synthetic, but also where it originates from, thereby fostering trust, accountability, and resilience against increasingly sophisticated synthetic speech attacks. Following the fusion paradigms introduced in Chapter 4(**MATA**), Chapter 5 (**MiO**), and Chapter 5(**SCAR**) this section also adheres to a consistent methodology. We begin by determining the most suitable PTMs for the downstream task and then formulate a tailored fusion strategy to achieve improved performance.

The main contributions are as follows:

- We present a comprehensive comparative study of SOTA speech PTMs that have demonstrated strong performance across prosodic-focused tasks, evaluating their capacity to capture discriminative prosodic signatures for SSSA.
- We demonstrate that representations from x-vector, a speaker recognition speech PTM, attains the best performance, which can be attributed to its speaker-centric pre-training that effectively captures prosodic characteristics.
- We propose **FINDER**, a novel framework that employs renyi divergence as a fusion mechanism to fuse heterogeneous PTM representations.

6.1.2 Modeling

Modeling with Individual PTM representations: We employ two downstream architectures with individual PTM representations: a FCN and a CNN. The FCN comprises three dense layers with 256, 128, and 64 neurons, followed by an output layer. The CNN architecture consists of two convolutional blocks, each built using a 1D convolutional layer and a max-pooling operation, followed by flattening and a stack of three dense layers identical to those in the FCN. Specifically, the first convolutional block contains 256 filters with a kernel size of 3, followed by batch normalization and max pooling with a pool size of 2. The second convolutional block employs 128 filters with a kernel size of 3, again followed by batch normalization and max pooling with a pool size of 2. The number of trainable parameters in the CNN models using individual PTM representations ranges from approximately 0.8M to 1.2M, depending on the dimensionality of the input PTM representations.

FINDER: Framework: We introduce **FINDER**, a novel framework designed for the effective fusion of PTM representations. The overall architecture is illustrated in Figure 6.1. Representations extracted from each PTM are first processed through two convolutional blocks that follow the same configuration as the CNN architecture used for modeling individual PTM representations. After the convolutional stages, the resulting feature maps are flattened and linearly projected to a 120-dimensional embedding space to ensure dimensional consistency across PTMs and to meet computational constraints. These projected features are then aligned using renyi divergence (RD). It is a statistical measure of dissimilarity between two probability distributions Van Erven and Harremos (2014). In our framework, RD is formulated as a loss function that quantifies the diver-

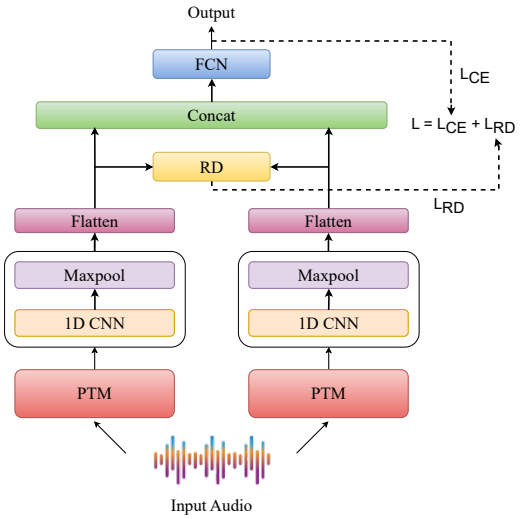


Figure 6.1: **FINDER**: RD refers to renyi divergence, L, L_{CE}, and L_{RD} correspond to the total loss, cross-entropy loss, and renyi divergence loss, respectively.

gence between the learned feature representations of two distinct PTMs. A lower RD value indicates stronger alignment between the representations. Let $\mathbf{z}_1$ and $\mathbf{z}_2$ denote the projected features obtained from two PTM branches. The RD ($\mathcal{L}_{\text{RD}}$) between their feature distributions is defined as:

$$\mathcal{L}_{\text{RD}} = \frac{1}{\beta - 1} \log \left(\sum_{j=1}^{K} (\mathbf{z}_{1,j} + \delta)^{\beta} (\mathbf{z}_{2,j} + \delta)^{1-\beta} \right), \tag{6.1}$$

where K denotes the embedding dimensionality, $\beta > 1$ controls the order of the divergence, and δ is a small constant introduced for numerical stability. The objective of this divergence term is to reduce the dissimilarity between heterogeneous PTM representations, encouraging them to occupy a shared, aligned feature space. For end-to-end training, the RD loss $\mathcal{L}_{\text{RD}}$ is jointly optimized with the cross-entropy loss $\mathcal{L}_{\text{CE}}$ used for the downstream classification task. The overall training objective is defined as:

$$\mathcal{L}_{\text{total}} = \eta \mathcal{L}_{\text{CE}} + (1 - \eta) \mathcal{L}_{\text{RD}}, \tag{6.2}$$

where η is a weighting hyperparameter that balances classification performance and representation alignment. The total number of trainable parameters ranges from approximately 1.3M to 1.5M, depending on the input representation dimension.

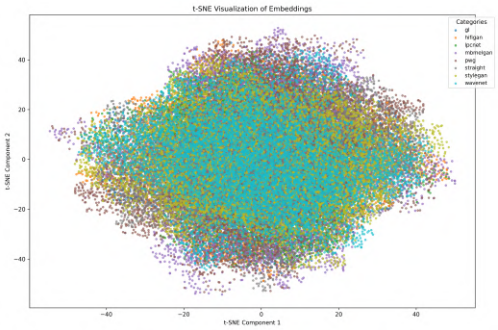
6.1.3 Experiments

Pre-Trained Representations: We use WavLM, Wav2vec2, XLS-R, and Whisper. In addition to the considered PTMs, we include x-vector Snyder *et al.* (2018*b*). Owing to its speaker-discriminative pre-training, x-vector has demonstrated strong performance across several prosodic tasks, including SER Chetia Phukan *et al.* (2023), shout intensity prediction Fukumori *et al.* (2023), and depression detection Egas-López *et al.* (2022), among others. We further consider Wav2Vec2-emo, a Wav2vec2-based model fine-tuned for SER. Since SER is inherently driven by prosodic cues, we hypothesize that the representations learned by Wav2vec2-emo may also be informative for SSSA.

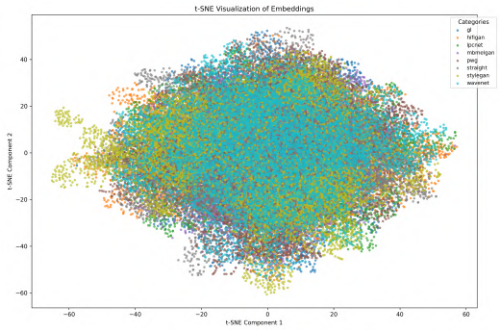
Benchmark Datasets: We conduct our experiments on two benchmark datasets, namely ASV and CFAD. For ASV, we merge the original training, validation, and test splits, resulting in 19 synthetic source classes (A01–A19). In contrast, CFAD contains 9 source classes (0–8). We adopt five-fold cross-validation for ASV and follow the official data split provided for CFAD. Additional details regarding both datasets are presented in Chapter 3.

Training Details: All models use a softmax activation function in the output layer to produce class probabilities. Training is performed using the Adam optimizer with a learning rate of 1e-3. Each model is trained for 40 epochs with a batch size of 32, and cross-entropy loss as the loss function. To mitigate overfitting, we apply dropout. For experiments involving **`FINDER`**, the $\mathcal{L}_{\text{RD}}$ parameters are set to $\alpha = 2$ and $\epsilon = 0.1$, with a loss-balancing weight of $\lambda = 0.4$. These hyperparameters are kept fixed across all experiments, as they consistently yielded strong performance during preliminary exploration.

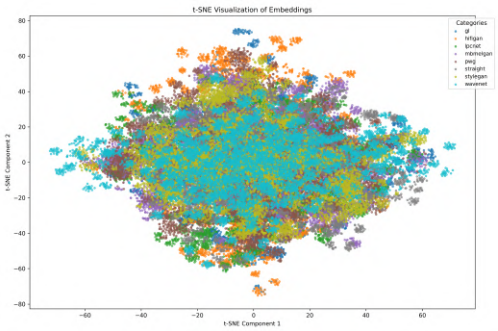
Experimental Results: We employ accuracy and EER as evaluation metrics, following prior work on SSSA Klein *et al.* (2024) and SSD Liu *et al.* (2023). For EER, we report the average one-vs-all scores. The performance of downstream models trained using individual PTM representations is summarized in Table 6.1. Across both datasets, x-vector consistently achieves the strongest performance, characterized by higher accuracy and lower EER. This advantage can be attributed to its speaker recognition pre-training, which inherently encourages the learning of rich prosodic representations—a trend that aligns with its effectiveness across multiple prosody-driven tasks such as



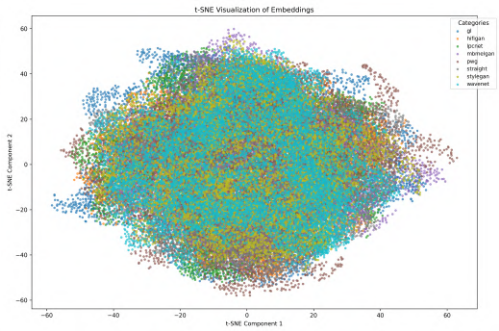
(a) WavLM



(b) Whisper

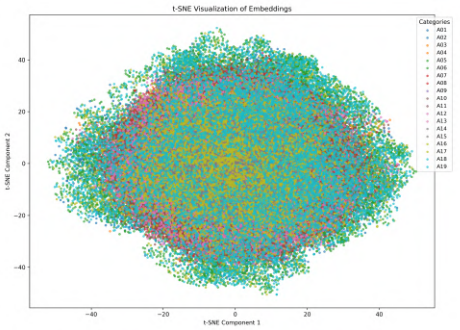


(c) x-vector

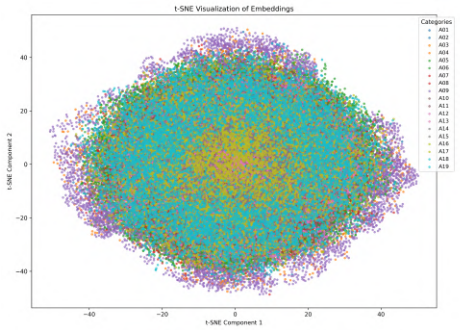


(d) Wav2vec2-emo

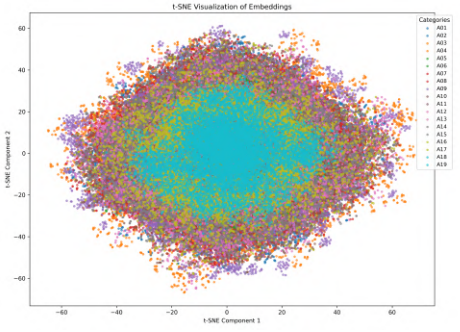
Figure 6.2: Visualization of PTMs Representational Space for CFAD



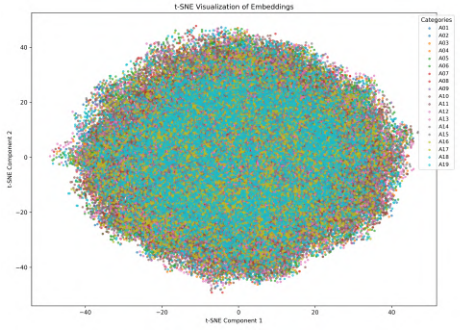
(a) WavLM



(b) Whisper



(c) x-vector



(d) Wav2vec2-emo

Figure 6.3: Visualization of PTMs Representational Space for ASV

Table 6.1: Performance comparison of individual PTM representations on ASV and CFAD. All results are reported in percentage (%), with ASV scores averaged over 5 folds. Higher accuracy and lower EER indicate better performance. Abbreviations: W2V—Wav2vec2, WLM—WavLM, XLR—XLS-R, WSR—Whisper, and XVC—x-vector. These abbreviations are kept same for Table 6.2.

Representations	**ASV**				**CFAD**			
	FCN		**CNN**		**FCN**		**CNN**	
	Accuracy	**EER**	**Accuracy**	**EER**	**Accuracy**	**EER**	**Accuracy**	**EER**
W2V	45.25	21.54	61.75	7.76	49.25	29.20	74.50	10.20
WLM	33.48	15.25	45.46	10.50	32.23	22.23	35.78	27.60
XLR	63.98	11.55	79.04	4.01	50.25	19.30	76.90	9.50
WSR	75.69	9.85	87.03	4.01	70.14	15.85	85.01	8.10
XVC	**87.48**	**4.42**	**97.60**	**2.03**	**74.58**	**10.02**	**91.39**	**4.40**
W2V-emo	78.45	8.40	86.35	2.50	65.30	12.12	86.50	8.60

Table 6.2: Performance comparison of fusion methods on ASV and CFAD. ASV results are averaged over five folds, with all metrics reported in percentage (%). Higher accuracy and lower EER indicate better performance.

Combinations	**ASV**				**CFAD**			
	Concatenation		**FINDER**		**Concatenation**		**FINDER**	
	Accuracy	**EER**	**Accuracy**	**EER**	**Accuracy**	**EER**	**Accuracy**	**EER**
W2V + W2V-emo	96.57	0.67	96.69	0.63	89.80	3.60	93.32	3.51
WLM + W2V-emo	93.60	1.10	94.60	1.08	85.65	5.70	88.03	5.55
XLR + W2V-emo	91.23	1.10	94.35	1.04	92.47	2.79	95.72	2.64
WSR + W2V-emo	96.03	0.66	96.66	0.63	94.95	1.50	98.47	1.41
XVC + W2V-emo	93.12	1.33	98.62	0.37	91.79	2.80	96.71	2.50
XVC + W2V	96.25	0.63	98.16	0.33	85.50	5.40	86.77	4.40
WLM + W2V	96.38	0.69	96.99	0.65	89.75	2.84	96.26	2.61
WSR + W2V	96.87	0.51	97.21	0.50	94.11	2.04	97.59	1.91
XLR + W2V	96.09	0.65	96.54	0.57	94.93	1.45	98.09	1.24
WLM + XLR	86.92	2.30	87.40	2.04	92.63	1.80	98.75	1.17
WSR + XLR	93.99	1.10	94.50	1.01	89.50	3.40	94.98	2.97
XVC + XLR	96.58	1.20	97.84	0.35	90.87	3.30	95.58	1.60
WSR + WLM	94.26	0.85	94.69	0.84	88.86	3.84	94.28	3.08
XVC + WLM	95.85	0.40	98.37	0.36	64.32	10.10	78.77	8.17
WSR + XVC	**97.16**	**0.32**	**98.91**	**0.26**	**95.00**	**1.10**	**99.01**	**1.07**

SER, shout intensity prediction, and depression detection Chetia Phukan *et al.* (2023); Fukumori *et al.* (2023); Egas-López *et al.* (2022). In contrast, monolingual PTMs such as Wav2Vec2 and WavLM exhibit inferior performance, likely due to their limited ability to capture source-specific prosodic characteristics. The behavior of Wav2Vec2-emo is consistent with expectations, as it is fine-tuned for SER; however, its performance remains inferior to that of x-vector. Whisper and XLS-R demonstrate improved performance compared to monolingual PTMs, reciprocating findings from our prior SSD study Phukan *et al.* (2024*a*) that multilingual PTMs are better at modeling diverse prosodic patterns, including variations in pitch, tone, and rhythm.

We further visualize the learned PTM representations using t-SNE plots, presented in Figures 6.2 and 6.3. These visualizations support the quantitative results, reveal-

ing more distinct clustering of source classes for x-vector. Finally, we observe that CNN-based downstream models consistently outperform their FCN counterparts, highlighting the effectiveness of convolutional architectures for modeling PTM representations. Table 6.2 reports the performance of PTM representation fusion using a baseline concatenation-based approach and the proposed **FINDER** framework. For the baseline method, we adopt the same architectural and training configuration as FINDER, with the exception of the RD loss. All other training settings are kept identical to ensure a fair comparison. Results indicate that fusing PTM representations using both fusion strategies consistently outperforms models based on individual PTM representations, highlighting the complementary nature of heterogeneous PTM representations. Notably, fusion via **FINDER** consistently surpasses the concatenation-based baseline across all settings, demonstrating the effectiveness of Divergence Guided Alignment. Among the evaluated combinations, Whisper and x-vector fused using **FINDER** achieve the best overall performance on both datasets. The strong performance of this PTM pair can be attributed to their complementary strengths. x-vector effectively captures prosodic cues due to its speaker-centric training objectives, while Whisper has demonstrated robustness in modeling prosodic characteristics relevant to SSD. Their combination therefore provides richer and more discriminative representations, resulting in superior performance in SSSA.

Dataset	Model	Accuracy(%)	EER(%)
ASV	**FINDER (WSR + XVC)**	**98.91**	**0.26**
	MiO(WSR + XVC)	97.75	0.68
	AASIST(W2V)	63.81	5.68
CFAD	**FINDER**	**99.01**	**1.07**
	MiO(WSR + XVC)	97.31	2.15
	AASIST(W2V)	77.92	9.69

Table 6.3: Comparison to prior SOTA works. Abbreviations: WSR—Whisper, XVC—x-vector, and W2V—Wav2vec2.

Comparison to prior SOTA: Since we aggregate all source classes across the training, validation, and test splits for ASV, direct comparison with prior studies is not feasible. Furthermore, to the best of our knowledge, this work represents the first investigation of SSSA on CFAD, and therefore no existing benchmarks are available for direct comparison. To facilitate a meaningful evaluation, we reimplement several SOTA methods originally proposed for SSSA and SSD and compare them against our approach. We adopt MiO Phukan *et al.* (2024*a*) (Chapter 5), our SOTA framework for SSD that combines multiple PTM representations, and implement it using Whisper and x-vector represen-

tations, which constitute the best-performing PTM pair in our experiments. For SSSA, we reimplement AASIST Jung *et al.* (2022) as the downstream model with Wav2Vec2 representations, following the setup described in Klein *et al.* 2024. Table 6.3 presents a comparative evaluation between the proposed method and these SOTA baselines. Our results show that **FINDER** consistently outperforms both of these methods, achieving SOTA performance and demonstrating its effectiveness for SSSA.

Statistical Significance: The statistical significance of the observed performance improvements was evaluated using paired McNemar's tests on identical test sets from ASV and CFAD. Pairwise comparisons between **FINDER** (Whisper + x-vector) and the strongest concatenation-based (Whisper + x-vector) baseline resulted in p-values of 0.00057 (ASV) and 0.00782 (CFAD), respectively. Since both values are below the conventional significance threshold of 0.05, the observed improvements can be considered statistically significant.

6.2 COFFE

In this section, we present the background and motivation for SSSA in the edge case when exposed to singing voice, emphasizing the heightened variability in pitch, timbre, and expressive dynamics that make synthetic singing voice particularly challenging than conventional synthetic speech. We present the background and motivation and establish the task of singing voice SSSA, which we term it as SVSSSA throughout this section for convenience. We then introduce **COFFE** Phukan *et al.* (2025*c*), a Divergence Guided Aligment framework that leverages chernoff distance, thereby enhancing performance in SVSSA.

6.2.1 Background and Motivation

"Imagine discovering a new song by your favorite artist, only to realize they never recorded it." With the rapid advancement of generative technologies, this scenario is no longer speculative. Synthetic singing voice have reached a level of realism where an artist's vocal timbre can be convincingly replicated, seamlessly blending speech articulation with musical tonality Zang *et al.* (2024*b*). While such developments expand

the boundaries of creative expression, they simultaneously raise serious concerns regarding authenticity, intellectual property protection, and the ethical use of AI in music generation. As synthesis techniques continue to advance, the challenge extends beyond simple detection—establishing the attribution of synthetic singing voice has become a critical requirement. Accurately tracing the origins of synthetic singing voice is essential for preserving artistic integrity and preventing misuse; however, this problem remains largely unexplored within the domain of synthetic speech forensics. While significant progress has been made in synthetic singing voice detection Zang *et al.* (2024*b*,*a*); Zhang *et al.* (2024*d*), the equally important problem of source attribution—identifying the specific model responsible for generating a deepfake—remains largely unexplored. In this work, we extend the concept of SSSA introduced in Chapter 6 (**FINDER**) to the domain of singing voice and formalize a new task, SVSSSA. Unlike conventional synthetic speech or singing voice detection, which focuses solely on classifying an audio sample as real or synthetic, SVSSSA aims to trace the origin of synthetic singing voice, thereby uncovering the underlying generative process. SSSA has attracted growing attention in recent years Yan *et al.* (2022*b*); Zhang *et al.* (2024*a*); however, analogous efforts for singing voice remain sparse. Early work by Müller et al. Müller *et al.* (2022*b*) employed RNN-based models to characterize signatures of seen and unseen synthetic speech sources. More recently, Klein et al. Klein *et al.* (2024) and Bhagtani et al. Bhagtani *et al.* (2024) demonstrated the effectiveness of SOTA speech PTMs, such as Wav2Vec2 and Whisper, for SSSA. These PTMs have since been widely adopted across related tasks, including our SSD work Phukan *et al.* (2024*a*) and synthetic singing voice detection Chen *et al.* (2024).

As the first study on SVSSSA, we explore a diverse set of PTMs and *hypothesize that multimodal PTMs, such as IB and LB, are particularly well suited for this task. Their cross-modal pretraining enables the learning of rich and complementary representations, making them effective at capturing subtle, source-specific characteristics in synthetic singing voice, including timbral cues, pitch variability, and synthesis artifacts*. To validate this hypothesis, we conduct a large-scale comparative evaluation involving multimodal PTMs, speech PTMs, and music PTMs. The inclusion of speech and music PTMs is motivated by prior synthetic singing voice detection studies that have demonstrated their effectiveness Zang *et al.* (2024*b*); Zhang *et al.* (2024*d*). Furthermore, inspired by our prior works in SSD Phukan *et al.* (2024*a*) and SSSA Chetia Phukan

et al. (2025) which show that fusing heterogeneous PTM representations can exploit complementary behaviors and yield improved performance, we also investigate PTMs fusion for SVSSSA. Additionally, improvement in performance due to fusion of PTM representations is also shown across tasks such as speech recognition Arunkumar *et al.* (2022), and synthetic singing voice detection Chen *et al.* (2024); Guragain *et al.* (2024). As such, to this end, we propose **COFFE** (Fusion via **C**hern**OFF** Distanc**E**), a novel framework that leverages chernoff distance (CD) as a loss function to align PTM representations within a shared feature space. Through the fusion of LB and IB using **COFFE**, we achieve the best overall performance, surpassing individual PTMs, baseline fusion strategies, and establishing new SOTA results on benchmark dataset for SVSSSA. Following the fusion paradigms introduced in Chapter 4 (**MATA**), Chapter 5 (**MiO** and **SCAR**), and Chapter 6(**FINDER**), this section adopts a consistent methodological framework. We first identify the most suitable PTMs for the downstream SVSSSA task and subsequently design a tailored fusion strategy to further enhance performance.

The key contributions are as follows:

- We introduce SVSSSA, a novel task that focuses on tracing the generative origins of synthetic singing voices extending conventional SSSA.
- We demonstrate that multimodal PTMs outperform unimodal speech and music PTMs for SVSSSA, owing to their multimodal pre-training.
- We propose **COFFE**, a novel fusion framework that leverages CD as an alignment loss. The fusion of LB and IB using **COFFE** yields the highest performance compared to individual PTMs and baseline fusion methods.
- We introduce the first comprehensive benchmark for SVSSSA.

6.2.2 Modeling

Modeling with Individual PTM representations: We employ two downstream architectures for modeling individual FM representations: FCN and CNN. The CNN architecture comprises of two 1D convolutional layers with 64 and 128 filters, respectively, each using a kernel size of 3 and followed by max-pooling layers with a pool size of 2. The resulting feature maps are flattened and fed into an FCN block consisting of a dense layer with 128 neurons. The final output layer contains 8 neurons with a softmax activation function, producing probability estimates over the source classes. The FCN-

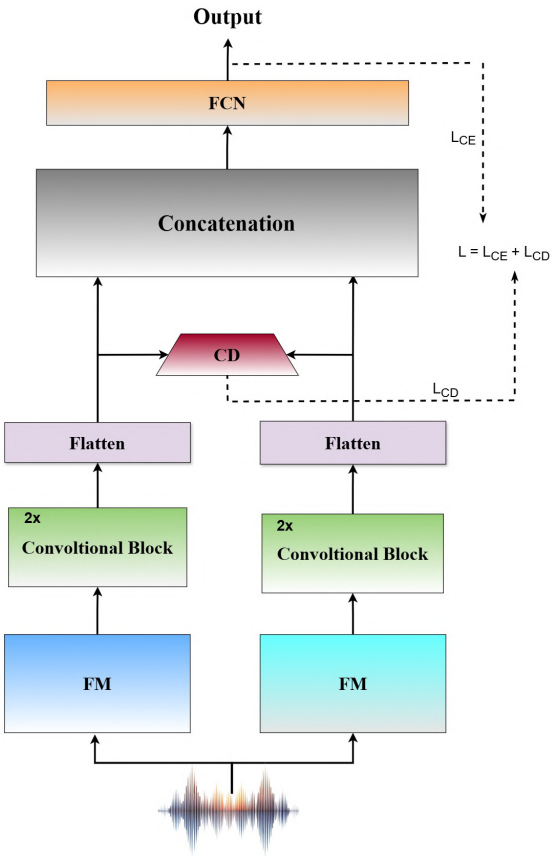


Figure 6.4: **COFEE**: FM denotes a PTM, and CD represents the chernoff distance. L_{CD}, L_{CE}, and L correspond to the chernoff distance loss, cross-entropy loss, and total loss, respectively.

based model follows the same architectural configuration as the FCN block used in the CNN.

COFFE: Framework: We propose **COFFE**, a novel framework for the fusion of representations from heterogeneous PTMs. The overall architecture is illustrated in Figure 6.4. Representations extracted from each PTM are first processed through two convolutional blocks identical to those used for modeling individual PTMs, each consisting of 1D convolutional layers followed by max-pooling operations. The resulting feature maps are then flattened. To align the representation spaces of heterogeneous PTMs, we employ CD Peng and Seetharaman (2012) as a novel divergence-based loss function. CD is particularly effective for minimizing the separability between feature distributions, thereby encouraging the PTM representations to converge toward a shared embedding space. Let $\mathbf{u}$ and $\mathbf{v}$ denote the feature distributions obtained from two PTM branches. The CD between these representations is defined as:

$$\mathcal{L}_{\text{CD}} = -\log\left(\sum k(\mathbf{u}_k)^{\rho}(\mathbf{v}_k)^{1-\rho}\right), \tag{6.3}$$

where $\rho \in (0, 1)$ controls the relative contribution of each PTM. A larger value of $\mathcal{L}_{\text{CD}}$ corresponds to greater separability between the representations; therefore, the objective is to minimize this loss to achieve effective alignment. The features are then concatenated and passed through a fully connected block comprising a dense layer with 128 neurons, followed by an output layer with 8 neurons corresponding to the synthetic singing voice source classes. The final training objective jointly optimizes the classification and alignment objectives and is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \gamma\mathcal{L}_{\text{CD}}, \tag{6.4}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss, $\mathcal{L}_{\text{CD}}$ is the CD loss, and γ is a weighting hyperparameter that balances the two objectives. The total number of trainable parameters in the proposed framework ranges from approximately 3M to 8M, depending on the dimensionality of the PTM representations.

6.2.3 Experiments

Pre-Trained Models: We employ LB and IB as multimodal PTMs. The speech PTMs include Unispeech-SAT, Wav2Vec2, WavLM, Whisper, XLS-R, x-vector, MMS, and HuBERT. For music PTMs, we use music2vec-v1, MERT-v1-95M, MERT-v0-public, MERT-v1-330M, and MERT-v0.

Dataset: We use CtrSVDD dataset for our experiments. Detail information is given in **Chapter 3**. The dataset has provided the official training, development, and evaluation partitions. As the training and development sets contain samples generated by the same systems (A01–A08), we utilize these splits for training and evaluating our models for SVSSSA.

Training and Hyperparameter Details: All models are trained for 50 epochs using the Adam optimizer with cross-entropy loss and a learning rate of 1e-3. Dropout is applied to mitigate overfitting. For experiments with **`COFFE`**, the hyperparameters s

PTM	FCN			CNN		
	Acc ↑	F1 ↑	EER ↓	Acc ↑	F1 ↑	EER ↓
UNI	45.56	42.22	22.25	48.76	46.56	18.95
W2V	59.17	52.12	17.12	64.03	57.90	13.66
WLM	35.22	31.26	27.87	38.96	36.91	20.18
XLR	74.90	64.03	10.85	77.51	69.73	8.73
WSR	67.92	64.69	12.32	72.65	65.11	10.70
MMS	76.36	67.56	8.97	80.41	76.83	7.31
XVC	63.48	61.17	13.04	66.52	63.95	12.32
LB	**79.69**	**77.26**	**6.98**	**82.37**	**79.80**	**5.35**
IB	78.50	74.41	7.54	81.92	77.90	6.19
Mv1	40.87	37.02	25.72	47.13	42.98	19.79
M95	47.95	45.12	21.44	50.48	47.85	18.71
Mpub	53.03	47.19	19.52	55.66	52.49	17.77
M330	58.64	54.93	15.95	67.96	63.88	12.19
Mv0	45.47	42.81	24.58	48.55	45.53	19.65

Table 6.4: Evaluation scores for different PTMs. Abbreviations: UNI—Unispeech-SAT, W2V—Wav2vec2, WLM—WavLM, XLR—XLS-R, WSR—Whisper, XVC—x-vector, Mv1—music2vec-v1, M95—MERT-v1-95M, Mpub—MERT-v0-public, M330—MERT-v1-330M, and Mv0—MERT-v0. All scores are reported in percentage (%). The same abbreviations are used in Tables 6.5 and 6.6.

and λ are set to 0.3 and 0.1, respectively, based on preliminary experiments.

Experimental Results: Although the dataset used in this study does not explicitly contain musical accompaniments, we include music PTMs in our evaluation, motivated by the intuition that their pre-training on musical data may enable them to implicitly capture rhythmic and temporal patterns inherent to singing voice. Such characteristics could be beneficial for SVSSSA. Table 6.4 reports the performance of downstream models trained using individual PTMs for SVSSSA. We evaluate the models using accuracy, F1-score, and EER. Accuracy is particularly well suited for source attribution tasks, as demonstrated by prior SSSA study Klein *et al.* (2024), including our own analysis in Chapter 6 (**`FINDER`**). We additionally report EER, which is a standard metric in synthetic singing voice detection and broader SSD tasks Zang *et al.* (2024*b*); Phukan *et al.* (2024*a*). For EER, we present average one-vs-all scores. The results show that multimodal PTMs consistently outperform both speech and music PTMs across all evaluation metrics. *These findings validate our hypothesis that multimodal PTMs are particularly effective for SVSSSA due to their cross-modal pre-training, which enables them to capture fine-grained, source-specific characteristics—such as timbral cues, pitch vari-*

	Concatenation			COFFE		
Combinations	**ACC ↑**	**F1 ↑**	**EER ↓**	**ACC ↑**	**F1 ↑**	**EER ↓**
UNI + W2V	61.78	54.02	17.13	62.36	61.96	15.20
UNI + WLM	52.33	44.28	23.11	54.96	53.61	19.63
UNI + XLR	74.60	61.24	11.97	72.31	78.50	7.39
UNI + WSR	68.49	62.67	13.04	73.18	71.37	9.62
UNI + MMS	73.23	68.49	9.16	73.31	78.22	7.94
UNI + XVC	65.89	58.65	11.80	67.34	66.34	8.61
UNI + LB	78.86	74.00	8.22	80.91	79.36	5.34
UNI + IB	72.10	57.77	12.65	76.36	74.36	10.37
UNI + Mv1	59.28	52.29	19.09	47.13	42.98	16.43
UNI + M95	63.64	57.26	16.69	68.36	67.63	11.91
UNI + Mpub	62.83	55.81	14.97	66.95	65.09	9.33
UNI + M330	66.74	60.54	13.63	71.69	70.34	11.53
UNI + Mv0	60.91	53.86	18.35	63.85	62.31	13.31
W2V + WLM	60.06	52.18	16.35	65.85	63.29	11.61
W2V + XLR	73.05	65.89	10.13	74.31	73.35	6.39
W2V + WSR	71.24	65.52	11.75	73.36	72.39	9.32
W2V + MMS	71.78	62.73	10.58	74.62	71.59	8.62
W2V + XVC	68.15	57.41	12.65	75.35	72.64	10.37
W2V + LB	77.02	72.36	9.54	81.36	80.91	6.32
W2V + IB	71.48	58.20	12.14	76.61	75.30	9.33
W2V + Mv1	60.23	50.59	17.23	62.93	61.39	13.96
W2V + M95	63.46	54.13	16.79	65.23	64.94	10.63
W2V + Mpub	62.51	52.15	16.52	64.36	63.96	11.91
W2V + M330	64.98	55.19	15.01	67.96	66.11	9.61
W2V + Mv0	61.38	51.19	17.13	63.31	62.96	12.33
WLM + XLR	73.00	59.27	11.06	75.66	74.31	6.97
WLM + WSR	67.81	60.31	12.87	69.34	68.36	8.63
WLM + MMS	75.87	66.99	10.34	76.93	77.23	10.01
WLM + XVC	65.41	58.69	12.16	66.31	64.96	8.19
WLM + LB	78.48	73.50	8.58	81.63	80.14	8.13
WLM + IB	69.41	59.61	9.98	78.54	70.13	8.64
WLM + Mv1	54.19	47.05	22.41	56.96	55.14	19.17
WLM + M95	58.18	50.63	18.84	61.08	60.31	16.42
WLM + Mpub	58.34	49.78	16.92	60.27	52.67	16.17
WLM + M330	63.42	57.45	14.54	65.63	64.49	11.16
WLM + Mv0	56.46	49.49	20.31	59.61	58.31	16.37
XLR + WSR	79.41	68.59	7.80	82.64	81.37	6.09
XLR + MMS	76.56	67.21	8.42	78.34	77.31	8.09
XLR + XVC	78.56	73.96	10.36	81.61	80.67	8.63
XLR + LB	77.80	72.40	8.08	79.38	77.31	8.31
XLR + IB	78.94	73.50	10.06	80.37	79.09	10.03
XLR + Mv1	74.07	66.53	11.16	74.93	73.52	8.34
XLR + M95	75.04	60.03	10.58	77.31	75.99	9.06
XLR + Mpub	76.11	66.09	9.61	78.05	74.70	8.20
XLR + M330	75.36	65.91	12.29	76.89	75.47	11.37
XLR + Mv0	71.36	61.11	11.04	72.89	71.67	9.53

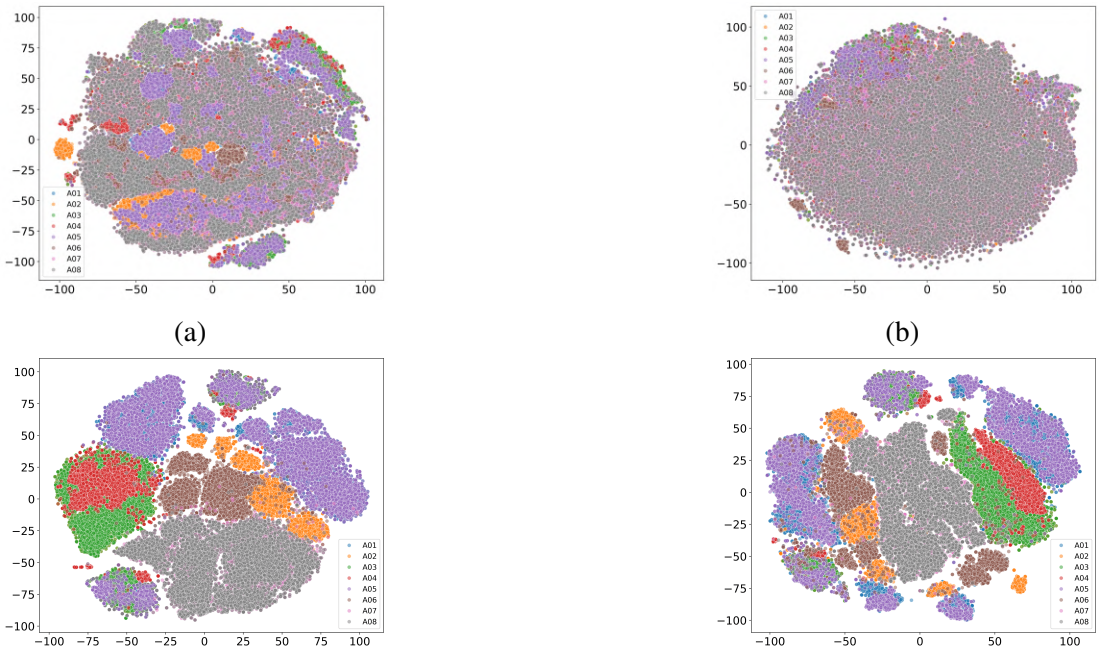
Table 6.5: Evaluation scores for different PTM combinations using feature concatenation and **COFFE**.

	Concatenation			COFFE		
Combinations	**ACC ↑**	**F1 ↑**	**EER ↓**	**ACC ↑**	**F1 ↑**	**EER ↓**
WSR + MMS	77.62	70.21	8.24	81.64	80.34	6.32
WSR + XVC	73.56	71.69	10.96	75.28	74.31	8.94
WSR + LB	78.64	74.07	8.36	81.39	80.31	6.17
WSR + IB	72.38	68.88	10.87	76.64	75.13	9.84
WSR + Mv1	69.78	64.32	12.75	71.18	70.61	9.49
WSR + M95	69.27	64.83	11.57	72.58	71.23	9.82
WSR + Mpub	70.48	65.90	12.08	72.71	69.57	11.80
WSR + M330	71.77	67.33	11.40	74.56	73.64	8.26
WSR + Mv0	69.71	64.42	12.72	73.64	72.54	10.67
MMS + XVC	76.54	71.35	10.31	77.96	76.68	8.22
MMS + LB	79.05	75.43	9.75	83.68	82.27	9.35
MMS + IB	76.59	71.38	10.06	78.34	76.13	9.08
MMS + Mv1	74.30	70.67	9.87	76.34	74.46	9.03
MMS + M95	72.98	63.75	9.57	75.38	74.49	9.64
MMS + Mpub	76.13	67.10	8.82	78.93	77.08	7.51
MMS + M330	75.56	70.77	8.98	77.34	76.31	7.34
MMS + Mv0	74.25	68.40	11.95	76.68	75.39	10.62
XVC + LB	82.02	80.15	6.54	83.64	82.66	4.67
XVC + IB	75.89	68.45	11.65	77.63	75.58	9.85
XVC + Mv1	65.23	63.56	14.63	67.73	66.69	11.63
XVC + M95	64.23	57.88	12.36	66.38	65.59	10.16
XVC + Mpub	61.35	57.46	11.36	63.28	61.03	9.34
XVC + M330	64.65	58.98	13.65	66.37	64.49	11.44
XVC + Mv0	59.63	55.41	11.65	62.38	60.34	10.74
LB + IB	**89.62**	**83.88**	**3.75**	**91.16**	**90.03**	**3.63**
LB + Mv1	77.60	72.30	8.91	78.63	77.62	8.03
LB + M95	77.70	77.20	8.65	80.37	78.86	7.39
LB + Mpub	77.63	72.88	9.20	79.78	76.62	8.33
LB + M330	80.40	77.59	7.63	82.29	81.17	7.39
LB + Mv0	79.06	75.84	8.04	82.23	81.18	7.93
IBD + Mv1	68.06	55.21	13.22	71.94	68.83	11.38
IB + M95	72.49	71.56	14.38	73.94	71.28	11.07
IB + Mpub	71.47	57.79	16.02	72.29	71.13	13.03
IB + M330	69.95	57.57	15.48	74.24	71.16	12.92
IB + Mv0	68.53	58.68	14.89	70.38	70.08	9.36
Mv1 + M95	57.94	51.52	19.77	61.93	58.39	12.83
Mv1 + Mpub	57.57	49.67	18.17	59.93	58.81	13.93
Mv1 + M330	62.89	57.46	13.25	63.49	63.02	11.52
Mv1 + Mv0	53.54	46.78	22.63	56.67	55.37	10.34
M95 + Mpub	59.77	53.01	19.51	62.64	59.89	16.51
M95 + M330	60.17	59.09	15.12	63.38	62.97	11.09
M95 + Mv0	57.39	51.06	19.84	59.96	58.64	13.81
Mpub + M330	63.20	56.66	14.60	66.33	64.85	9.57
Mpub + Mv0	57.65	49.43	18.30	59.29	56.73	13.38
M330 + Mv0	61.06	54.20	16.02	64.34	63.08	9.38

Table 6.6: Evaluation scores for different PTM combinations using feature concatenation and **COFFE**. All metrics are reported in percentage (%). This Table is continuation of Table 6.5, as due to space constraint, it couldn't be fitted within single page limit.

ability, and synthesis artifacts. Among the multimodal PTMs, LB and IB achieve the strongest performance with both FCN and CNN downstream architectures. Overall, CNN-based models consistently outperform their FCN counterparts across most PTMs. Following multimodal PTMs, multilingual speech PTMs demonstrate the second-best performance. This behavior can be attributed to their exposure to diverse language speech data during pre-training, which is particularly relevant given that the dataset used in this study contains singing voices in Chinese and Japanese. In contrast, monolingual speech PTMs—such as WavLM, Unispeech-SAT, and Wav2Vec2—exhibit inferior performance, likely due to the mismatch between their pre-training languages and the downstream data distribution. Music PTMs report the lowest performance overall, indicating limited effectiveness in capturing source-specific characteristics required for SVSSSA. An interesting observation is the strong performance of x-vector, a comparatively lightweight speech PTM trained for speaker recognition. Despite its smaller model size, x-vector outperforms both monolingual speech and music PTMs, suggesting that its speaker-discriminative pre-training objective equips it with an enhanced ability to model source-specific traits relevant for attribution. To further analyze the learned representations, we visualize the raw PTM representations using t-SNE, as shown in Figure 6.5. These visualizations reveal more distinct clustering of source classes for multimodal PTMs, providing qualitative support for our hypothesis and supporting the quantitative results.

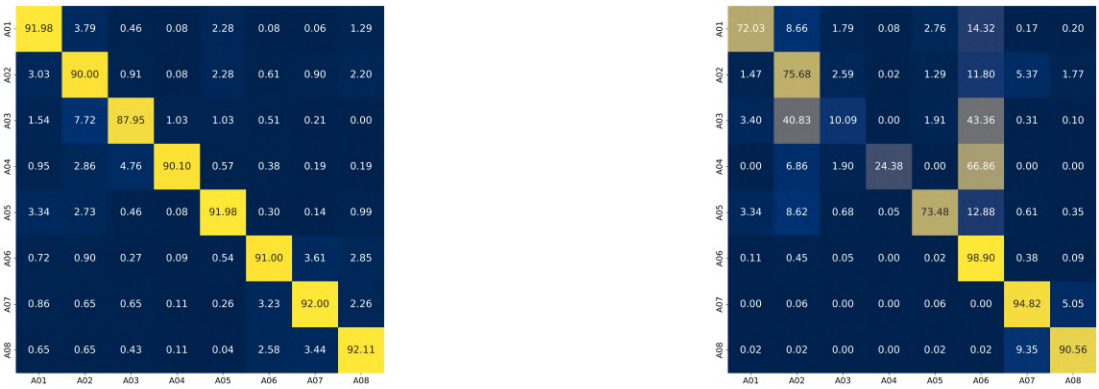
Table 6.5 and 6.6 summarizes the evaluation results obtained using different combinations of PTMs. We adopt a concatenation-based fusion strategy as the baseline. To ensure a fair comparison, the baseline follows the same architectural design and training configuration as **COFFE**, with the only difference being the exclusion of the CD loss. Overall, the results indicate that PTM fusion using **COFFE** consistently outperforms the baseline concatenation-based fusion approach. In particular, fusing the multimodal PTMs LB and IB using **COFFE** yields the best performance across all evaluated PTM combinations, as well as outperforming models based on individual PTMs. This highlights the ability of **COFFE** to effectively exploit the complementary strengths of multimodal PTMs. Furthermore, the confusion matrices shown in Figure 6.6 provide additional qualitative evidence of this improvement, illustrating enhanced classification accuracy when using **COFFE** with LB–IB fusion, compared to using LB alone with a CNN downstream. Finally, the results presented in this work establish a strong bench-



(a) (b)

(c) (d)

Figure 6.5: t-SNE visualizations- (a) XLS-R (b) MERT-v1-330M (c) LB (d) IB



(a) (b)

Figure 6.6: Confusion matrices: (a) **COFFE (LB + IB)**, (b) CNN (LB). The x-axis shows predicted values, and the y-axis shows true values.

mark for future research on SVSSSA.

Statistical Significance: The statistical significance of the reported improvements was assessed using paired McNemar's tests on identical test set. Comparing **COFFE** (LB + IB) against the strongest concatenation-based ((LB + IB) baseline yielded p-value of 0.00298 indicating that the observed gains are statistically significant at the 0.05 level.

6.2.4 Chapter Summary

This chapter examined representation fusion from the perspective of statistical alignment of PTMs internal representation distributions, in contrast to the geometric and interaction-based paradigms explored in earlier chapters. Through RD (**FINDER**) and CD (**COFFE**), we demonstrated that explicitly aligning the distributions of heterogeneous PTMs representational space leads to substantial gains in SSSA including the scenario of synthetic singing voice (SVSSSA). By encouraging representations to converge in a shared statistical space, Divergence Guided Alignment provides a flexible, principled mechanism for enabling PTMs to speak a common statistical language. Viewed alongside Chapter 4 and Chapter 5, this chapter completes a cohesive triad of fusion principles—geometry, interaction, and divergence. Each paradigm offers a distinct and complementary starting point for achieving proximal interpolative alignment, collectively forming a unified foundation for designing effective fusion strategies for improved affective and secure VIS.

CHAPTER 7

Hybrid Alignment: Combination of Interaction Guided with Geometry Guided alignment

Chapter 5 demonstrated that Interaction Guided Alignment mechanisms such as bilinear pooling and nested cross-attention—are effective for modeling complementary relationships between heterogeneous PTM representations. Separately, Geometry Guided Alignment methods have proven valuable by embedding PTM representations into structured geometric spaces. While interaction mechanisms capture fine-grained, instance-level dependencies between representations, geometric alignment enforces global structural consistency across embedding spaces; their combination enables both local relational modeling and global representational harmonization, leading to improved downstream performance. Building on insights from these earlier chapters, this Chapter introduces **PARROT** Phukan *et al.* (2025*b*), a Hybrid Alignment framework for SER that unifies Interaction Guided Alignment with Geometry Guided Alignment. While prior chapters has primarily focused on fusing transformer-based PTMs representations, our exploration extends this space by integrating representations from mamba-based PTMs with transformer-based PTMs, uncovering complementary behaviors arising from their distinct architectural differences. We begin by presenting the background and motivation for fusing PTMs of varied architectures, followed by a detailed explanation of the **PARROT** framework for more effective SER.

7.1 PARROT

7.1.1 Background and Motivation

From the beginning of this decade, the field of SER underwent a major shift with the rise and widespread accessibility of self-supervised learning PTMs wen Yang *et al.* (2021). These models, trained on vast and diverse speech corpora, not only deliver substantial performance improvements but also eliminate the need to build systems entirely from scratch. Their introduction has accelerated progress in SER, motivating researchers to

examine a variety of SOTA PTMs Yang (2023); Atmaja and Sasou (2022); Osman *et al.* (2024). For example, Pepino et al. Pepino *et al.* (2021) demonstrated the effectiveness of Wav2vec2 paired with LSTM and CNN classifiers, outperforming traditional feature sets such as eGeMAPS and spectrogram-based inputs. Similarly, Morais et al. Morais *et al.* (2022) provided an extensive comparison of PTMs like Wav2vec2 and HuBERT combined with different downstream networks. Notably, most of these PTMs are built on transformer-driven architectures, which are particularly strong at modeling contextual cues in speech signal.

More recently, a different class of PTMs has gained momentum within the speech community: mamba-based PTMsYadav and Tan (2024); Erol *et al.* (2024); Shams *et al.* (2024). Built on the structured state-space model (SSM) architecture, mamba offers an appealing alternative to transformers by capturing long-range dependencies with linear computational complexity Gu and Dao (2023). Early studies show that PTMs derived from this architecture achieve strong results on a range of sequence modeling tasks, and emerging evidence indicates that they can match—or in some cases outperform—their transformer-based counterparts for SER Yadav and Tan (2024). In parallel, earlier research has shown that combining PTMs representations has been advantageous for SER Wu *et al.* (2023*a*), likely because different PTMs encode complementary aspects of the speech signal. Similar benefits of fusion have also been observed in related domains such as automatic speech recognition Arunkumar *et al.* (2022) and SSD Phukan *et al.* (2024*a*). However, existing fusion strategies have almost exclusively focused on transformer-based PTMs.

In this work, we explore the fusion of mamba-based and transformer-based PTMs for SER, motivated by the idea that these architectures capture complementary aspects of the speech signal. *Specifically, we hypothesize that combining them can produce richer and more resilient representations: transformers are well suited for modeling complex global dependencies, whereas mamba-based models handle long-range patterns with computational efficiency.* To the best of our knowledge, this is the first attempt to fuse such architecturally diverse PTMs for SER. To this end, we introduce **PARROT** (**PAR**allel B**R**anch Hadamard **O**ptimal **T**ransport), a novel fusion framework designed for fusing these heterogeneous PTM representations. **PARROT** uses a dual-branch structure: one branch applies the hadamard product to capture fine-grained, element-wise interactions Interaction Guided Alignment) between the two PTMs, while

the second branch employs optimal transport to align their feature distributions at a global level (Geometry Guided Alignment). Together, these mechanisms yield a fused representation that is both locally sensitive and globally coherent, ultimately enhancing SER performance.

The main contributions are:

- We present **PARROT**, a novel fusion framework built on a dual-branch design that combines a hadamard product with optimal-transport, enabling it to jointly capture fine-grained local relationships and broader global structure for improved SER performance.
- Using **PARROT** with combination of mamba- and transformer-based PTMs delivers the topmost results across multiple SER benchmarks—CREMA-D (*English*), Emo-DB (*German*), and MESD (*Mexican Spanish*). It surpasses the performance of each individual PTM as well as the fusion of only transformer-based PTMs and baseline fusion techniques. Since the individual PTMs involved already represent SOTA for SER, the gains achieved by **PARROT** establish it as new SOTA.

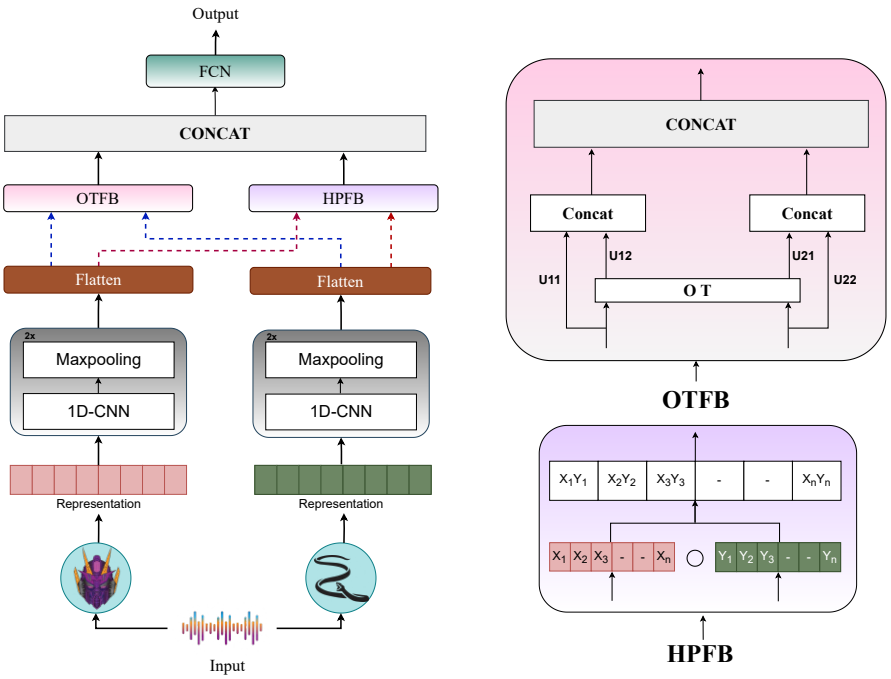


Figure 7.1: Proposed framework: **PARROT**. OTFB and HPFB denote the Optimal Transport Fusion Block and Hadamard Product Fusion Block, respectively. U_{11} and U_{22} represent the feature spaces of PTM 1 and PTM 2, while U_{12} and U_{21} correspond to the transported feature spaces from PTM 2 to PTM 1 and from PTM 1 to PTM 2, respectively.

7.1.2 PARROT: Framework

We propose **PARROT**, a novel framework for integrating heterogeneous PTM representations. The overall architecture is illustrated in Figure 7.1. The representations extracted from the two PTMs are first passed through identical stacks of 1D convolutional blocks (using the same filter settings as in the Experiments: Modeling with Individual PTM representations below). After flattening, both outputs are linearly projected into a 120-dimensional latent space. **PARROT** introduces a dual-branch fusion module that jointly models local interactions using hadamard product and global alignment using optimal transport between the PTMs. Let the projected representations be denoted by $\mathbf{A}$ and $\mathbf{B}$. To capture fine-grained local interactions, we compute an element-wise interaction: $\mathbf{H} = \mathbf{A} \odot \mathbf{B}$ where $\odot$ denotes hadamard product. Although this preserves detailed structural relationships, it does not guarantee global compatibility between $\mathbf{A}$ and $\mathbf{B}$. To capture the global interaction, optimal transport comes to picture. Firstly, a normalized euclidean cost matrix $\mathbf{D}$ is constructed as:

$$\mathbf{D} = \frac{\|\mathbf{A} - \mathbf{B}\|_2}{\max(\|\mathbf{A} - \mathbf{B}\|_2)}.$$

The transport plan $\mathbf{T}$ is then computed using the sinkhorn algorithm: $\mathbf{T} = \text{Sinkhorn}(\mathbf{D})$. Using $\mathbf{T}$, each representation is transported into the other's geometric structure:

$$\mathbf{A}' = \mathbf{T}\mathbf{A}, \qquad \mathbf{B}' = \mathbf{T}^\top \mathbf{B}.$$

The transported and original features are concatenated to form:

$$\mathbf{U}_A = \text{Concat}(\mathbf{A}', \mathbf{A}), \qquad \mathbf{U}_B = \text{Concat}(\mathbf{B}', \mathbf{B})$$

The outputs are merged into the final fused feature:

$$\mathbf{U} = \text{Concat}(\mathbf{U}_A, \mathbf{U}_B)$$

Finally, the outputs of both the parallel branches are concatenated and passed through a FCN with a 128-unit dense layer followed by the output layer with softmax activation

function. This design allows **PARROT** to retain local structure while ensuring global distributional alignment across PTMs. Depending on the representation-dimension of PTMs pair used, the number of trainable parameters in **PARROT** ranges from approximately 3.2M to 13M.

PTMs	CREMA-D		Emo-DB		MESD	
	Acc	F1	Acc	F1	Acc	F1
SVM						
A(T)	60.96	59.85	68.96	68.10	70.96	69.63
A(S)	68.96	67.10	78.96	77.65	71.63	70.85
A(B)	66.99	65.25	81.65	80.94	75.69	74.62
WLM	61.92	60.36	85.09	84.63	45.96	44.12
HBT	64.25	63.98	86.16	85.31	60.94	59.76
W2V	59.86	58.23	86.31	85.93	61.37	60.88
UNI	62.93	61.09	77.69	76.38	33.95	32.16
MMS	66.94	65.28	70.11	69.89	81.03	80.07
FCN						
A(T)	62.96	61.85	70.96	69.36	72.96	71.85
A(S)	69.52	68.20	80.66	79.63	76.93	75.11
A(B)	68.41	67.11	83.63	82.89	77.92	76.13
WLM	62.96	61.08	87.85	86.36	47.98	46.21
HBT	66.29	65.85	87.33	86.21	61.20	60.96
W2V	61.01	60.36	88.63	87.03	62.78	61.96
UNI	64.82	63.33	79.65	78.51	36.96	35.21
MMS	67.52	66.36	71.39	70.88	82.66	81.39
CNN						
A(T)	63.60	63.60	71.03	70.14	73.99	73.93
A(S)	69.51	69.49	81.31	79.27	78.61	78.53
A(B)	69.91	69.90	84.11	82.90	78.03	77.96
WLM	67.96	66.25	89.14	88.69	48.55	48.16
HBT	69.95	68.23	88.26	87.11	62.43	62.43
W2V	61.18	60.88	90.01	89.62	63.01	63.24
UNI	65.28	64.11	81.36	80.96	38.15	37.91
MMS	68.25	67.96	72.90	66.24	83.24	83.10

Table 7.1: Evaluation scores on CREMA-D, Emo-DB, and MESD. All scores are reported in percentage (%) and averaged over five folds. Abbreviations: Audio-Mamba variants—Tiny (A(T)), Small (A(S)), and Base (A(B)); WLM—WavLM; HBT—HuBERT; W2V—Wav2vec2; UNI—Unispeech-SAT. Acc and F1 denote Accuracy and macro-average F1, respectively. The same abbreviation scheme used in Table 7.1 is consistently maintained in Table 7.2.

7.1.3 Experiments

Modeling with Individual PTM Representations: We outline the downstream modeling architectures used with the individual PTMs, as well as those integrated within the proposed alignment framework, **PARROT**. For standalone PTM evaluation, we employ three classifier types: SVM, FCN, and CNN. The SVM is used with its default parame-

	CREMA-D				Emo-DB				MESD			
	Concatenation		PARROT		Concatenation		PARROT		Concatenation		PARROT	
Fusion	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
A(T)+WLM	63.37	62.89	64.38	63.38	76.96	75.61	78.69	77.25	44.38	43.81	46.85	45.28
A(T)+HBT	65.54	64.86	66.82	65.76	76.28	75.58	77.62	76.28	44.85	43.38	45.28	44.39
A(T)+W2V	60.03	59.94	61.76	60.58	75.14	74.61	76.93	75.58	45.03	44.97	46.92	45.22
A(T)+UNI	62.56	61.14	63.94	62.28	74.94	73.39	75.58	75.16	44.34	43.08	45.34	44.28
A(T)+MMS	62.95	61.59	63.86	62.47	76.85	75.28	77.78	76.94	45.37	44.65	46.39	45.80
A(S)+WLM	63.48	62.94	64.45	63.94	77.26	76.18	79.54	78.51	46.88	45.82	47.58	46.64
A(S)+HBT	62.14	61.19	63.68	62.24	77.36	76.52	78.82	77.34	45.28	44.28	46.62	45.82
A(S)+W2V	61.94	60.68	62.34	61.58	76.95	75.28	77.98	76.25	46.82	45.25	47.68	46.62
A(S)+UNI	61.64	60.17	62.84	61.34	75.82	74.36	76.52	75.94	44.29	43.57	46.94	45.82
A(S)+MMS	63.13	62.29	64.86	63.34	74.22	73.58	75.36	74.15	46.82	45.28	48.96	47.28
A(B)+WLM	63.96	62.10	67.56	67.44	79.41	78.64	80.37	78.56	47.31	46.85	49.13	48.87
A(B)+HBT	**70.63**	**69.79**	**73.68**	**72.90**	**89.92**	**88.24**	**92.24**	**91.53**	58.31	56.08	59.54	58.94
A(B)+W2V	69.96	69.21	71.98	70.94	88.94	87.64	89.44	88.57	56.37	55.39	57.23	56.64
A(B)+UNI	61.27	60.96	62.53	62.75	73.46	72.68	74.77	73.17	37.91	36.49	38.15	36.77
A(B)+MMS	61.25	60.97	63.26	63.24	65.96	64.05	66.36	57.29	**66.37**	**65.28**	**69.05**	**68.72**
WLM+HBT	68.99	67.64	69.91	69.85	80.96	79.64	81.31	79.90	53.94	52.17	54.34	54.12
WLM+W2V	67.96	66.21	68.30	68.23	86.94	85.61	87.85	87.70	54.96	53.29	55.49	55.14
WLM+UNI	66.93	65.07	67.90	67.71	76.68	75.61	77.57	77.52	48.64	47.34	49.71	49.59
WLM+MMS	65.39	64.07	67.90	67.84	87.96	86.09	88.29	87.54	53.94	52.18	54.34	54.20
HBT+W2V	69.37	68.13	70.52	69.32	77.91	76.63	78.50	76.97	53.94	52.28	54.34	53.47
HBT+UNI	69.58	68.37	70.61	69.51	73.49	72.94	74.77	74.13	57.49	56.19	58.96	58.96
HBT+MMS	69.37	68.96	71.52	70.22	87.91	86.34	88.69	87.33	64.94	63.28	65.32	64.74
W2V+UNI	67.94	66.39	68.96	67.16	75.39	74.17	76.64	76.37	57.19	56.41	58.38	58.51
W2V+MMS	64.19	63.61	65.21	65.01	80.91	79.31	81.52	80.96	72.14	71.39	71.10	70.96
UNI+MMS	68.34	67.36	69.63	67.45	73.91	72.33	74.63	73.93	49.68	48.31	50.29	49.50

Table 7.2: PTMs combinations pairs evaluation scores for CREMA-D, Emo-DB, and MESD. All scores are reported in percentage (%) and averaged over five folds.

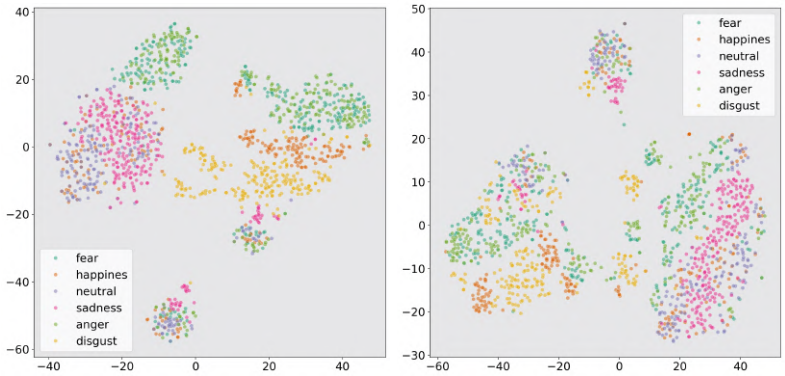
ter configuration. The CNN-based downstream consists of two 1D convolutional layers with 64 and 128 filters, respectively, each using a kernel size of 3 and ReLU activation. Each convolutional layer is followed by a max-pooling operation. The resulting feature map is flattened and fed into an FCN comprising a dense layer with 128 ReLU-activated units, followed by a softmax layer for multi-class classification. The FCN-only baseline adopts the same dense architecture as the FCN block used in the CNN setup.

Pre-Trained Models: We use Audio-Mamba variants — Tiny, Small, and Base; WavLM-Base; HuBERT; Wav2vec2; Unispeech-SAT; and MMS. The details of the PTMs are presented in Chapter 2.

Dataset: We use CREMA-D, Emo-DB and MESD datasets in our experiments. The details are given in Chapter 3.

Experimental Results: Table 7.1 summarizes the performance of each PTM when paired with different downstream models. Across datasets, CNN-based models generally deliver the strongest results, surpassing both SVM and FCN counterparts in accuracy and macro-F1. FCNs consistently outperform SVMs, further reinforcing the ad-

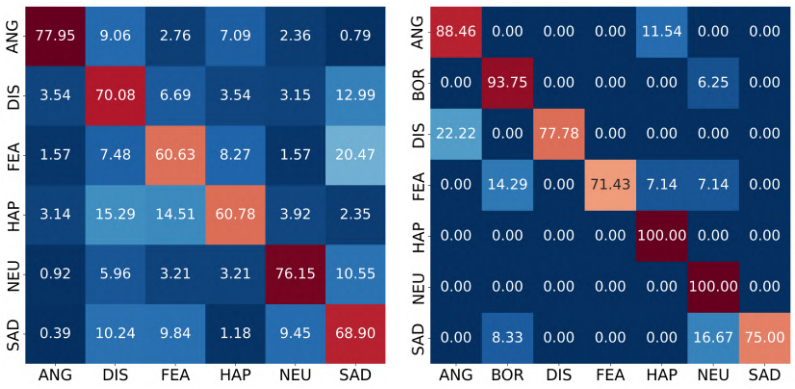
vantage of neural downstreams for modeling PTM representations. Within the mamba family, the base model emerges as the most reliable performer, outperforming the tiny and small variants on every dataset. Its gains can be attributed to its larger model size, which allows it to encode richer information crucial for SER. We also observe that the choice of downstream network noticeably affects performance, an effect previously noted in the literature Zaiem *et al.* (2023). For transformer-based PTMs, the results are more dataset-dependent. Some PTMs excel on particular corpora while lagging on others, underscoring the impact of dataset characteristics on downstream SER effectiveness. MMS achieves the strongest results on MESD, likely due to its multilingual pre-training. However, this advantage does not transfer uniformly; despite its broad linguistic coverage, MMS performs among the weakest models on Emo-DB. Similarly, most monolingual transformer PTMs underperform on MESD yet show competitive or superior results relative to mamba PTMs on Emo-DB. Taken together, the findings reveal that neither mamba-based nor transformer-based PTMs consistently dominate across datasets.



(a) (b)

Figure 7.2: t-SNE visualizations for CREMA-D: (a) **PARROT** with Audio-MAMBA (base) and HuBERT; (b) **PARROT** with Audio-MAMBA (base) and Wav2vec2

Table 7.2 summarizes the performance of various combinations of PTMs represenations. As a baseline, we use a simple concatenation-based fusion approach, implemented by removing the hadamard and optimal transport branches from the **PARROT** architecture (Figure 7.1) while keeping all other components and training settings identical for a fair comparison. Across all datasets, **PARROT** consistently yields higher accuracy and F1 scores than this concatenation baseline, demonstrating the advantage of its hybrid interaction–geometry fusion mechanism. Moreover, nearly all PTM combinations fused with **PARROT** outperform their respective individual PTMs, whereas



(a) (b)

Figure 7.3: Confusion matrices for **PARROT** with Audio-MAMBA (base) and HuBERT: (a) CREMA-D, (b) EMO-DB. The x-axis and y-axis represent predicted and true labels, respectivel

concatenation-based fusion often struggles to surpass single-PTM performance. Only in a few cases—primarily when mamba and transformer-based PTMs naturally complement one another—does simple concatenation offer marginal gains. In contrast, **PARROT** brings out this complementary behavior much more effectively. For instance, the fusion of Audio-Mamba (base) with HuBERT yields the highest scores on both CREMA-D and Emo-DB, while combining Audio-Mamba (base) with MMS delivers the best results on MESD. *These findings empirically validate our central hypothesis: heterogeneous fusion of mamba and transformer-based PTMs produces richer and more robust representations for SER, with transformers capturing global contextual structure and mamba models excelling at efficient long-range temporal modeling*. Overall, **PARROT** not only surpasses the best-performing individual PTMs—which previously represented the SOTA in SER wen Yang *et al.* (2021); Yadav and Tan (2024)—but also outperforms most transformer–transformer PTMs fusion and the concatenation baseline, establishing its effectiveness as new SOTA for SER. Figure 7.2 presents t-SNE visualizations of penultimate-layer representations, and Figure 7.3 shows the confusion matrices for the Audio-Mamba (base) + HuBERT combination through **PARROT**.

Statistical Significance: To assess the statistical reliability of the observed improvements, paired McNemar's tests were conducted using predictions from the same test sets across CREMA-D, Emo-DB, and MESD. The resulting p-values were obtained through pairwise comparisons between the proposed framework (Audio Mamba (Base) + HuBERT fused via **PARROT**) and the strongest competing baseline (Audio Mamba (Base) + HuBERT fused via Concatenation), with all values remaining below the com-

monly adopted significance threshold of 0.05 (0.00419 (CREMA-D), 0.00291 (Emo-DB), 0.00323 (MESD)).

7.2 Chapter Summary

This chapter brought together ideas from the previous fusion families to propose **PARROT**, a hybrid framework that combines hadamard product Interaction Guided Alignment) with optimal-transport (Geometry Guided Alignment). **PARROT** demonstrates that interaction and geometry are not competing paradigms but complementary components that, when integrated, can capture deeper representational relationships for improved affective and secure VIS.

CHAPTER 8

Conclusions, Future Work and Publications

8.1 Conclusions

We address two challenges of VIS in this thesis: enabling machines to robustly understand human affect and equipping them with the ability to detect and reason about synthetic speech. Recent advances in large-scale speech PTMs have significantly enhanced performance in both areas. However, these models are inherently heterogeneous, differing widely in architectural design, training objectives, linguistic coverage, and modality grounding. As a result, each model encodes a distinct—and often incomplete—view of the speech signal. As demonstrated throughout this thesis, no single PTM consistently outperforms others across datasets, languages, acoustic conditions, or adversarial settings. This observation leads to the central question of the work: rather than selecting or scaling individual PTMs, how can their diversity be systematically leveraged to build more effective affective and secure VIS.

To answer this question, this thesis introduced Proximal Interpolative Fusion as a unifying conceptual and methodological framework. At its core, Proximal Interpolative Fusion reframes fusion not as a post-hoc aggregation step, but as a structured process that first reduces incompatibility between heterogeneous PTM representations and then interpolates them into a shared yet expressive fused space. The notion of *proximity* emphasizes aligning PTM representations in task-relevant geometric, statistical, or interaction spaces. In contrast, *interpolation* ensures that fusion preserves complementary information instead of collapsing diverse inductive biases into a single dominant representation. This perspective departs fundamentally from naive concatenation or ensembling and forms the backbone of all fusion strategies proposed in this thesis.

Building on this paradigm, the chapters of the thesis systematically explored three complementary alignment families, followed by a hybrid formulation that unifies their strengths. Firstly, Geometry Guided Alignment, we proposed, **`HYFuse`**, **`MATA`**, **`MERLINN`**

that have demonstrated that embedding heterogeneous PTM representations into structured geometric spaces enables effective fusion when hierarchical structure or distributional mismatch is prominent. These approaches showed particular strength in SER, NVER, and special case of SSD when exposed to children's voice, highlighting the importance of respecting the latent geometry of representations prior to fusion. Secondly, within Interaction Guided Alignment, we propose, **MiO**, **SCAR** which shows that modeling explicit cross-representation interactions—via bilinear pooling or nested cross-attention—are crucial for modeling fine-grained dependencies between PTMs, especially in SSD scenarios. Complementing these approaches, Divergence Guided Alignment family (**FINDER**, **COFFE**) established that harmonizing the distributional behavior of PTMs provides a principled mechanism for SSSA tasks.

The final chapter, centered around Hybrid Alignment proposes **PARROT** which takes insights from the previous chapters and demonstrating that Interaction Guided Alignment and Geometry Guided Alignment are not competing paradigms but mutually reinforcing ones. By combining a hadamard product with optimal transport, **PARROT** illustrated how architectural heterogeneity—specifically between transformer-based and mamba-based PTMs—can be transformed from a liability into a source of strength for improved SER.

In summary, this thesis demonstrates that embracing PTM heterogeneity through principled fusion offers a powerful pathway toward more improved affective and secure VIS. By formalizing Proximal Interpolative Fusion paradigm, this work provides both conceptual clarity and practical tools for the next generation of affective and secure VIS. However, a practical consideration arising from this thesis is the computational overhead associated with fusion of PTMs. While combining multiple PTMs consistently improves performance and robustness, which remains the primary focus of this thesis, it also increases model complexity compared to single-PTM VIS. Consequently, practical VIS deployments must carefully balance the trade-off between accuracy, robustness, and computational efficiency. Future directions towards improving the deployability of VIS built on PTMs representations fusion are discussed in Future Works below.

Furthermore, in a practical deployment setting, the affective and secure dimensions of VIS need not operate as independent components and can instead complement one another within a unified pipeline. For instance, in a mental healthcare application, in-

coming speech may first pass through secure VIS modules that verify authenticity and identify potential generation sources before being processed by affective VIS modules for emotion understanding and behavioral analysis. Such a pipeline ensures that downstream affective reasoning is performed only on trusted speech inputs, thereby improving the reliability, robustness, and trustworthiness of the overall system.

Takeaway from the thesis: Beyond the individual frameworks proposed in this thesis, an important outcome of this research is the realization that no single alignment strategy is universally optimal across all affective and secure VIS tasks. Instead, the effectiveness of an alignment methodology depends largely on the nature of the heterogeneity present between the PTMs being fused, as well as the characteristics of the downstream task. Through the investigations presented in this thesis, it became evident that Geometry Guided Alignment is particularly effective when the primary challenge arises from differences in the latent manifold structure of the representations, often stemming from variations in model architectures, pre-training objectives, or learned information. In contrast, Interaction Guided Alignment is more suitable when success depends on capturing fine-grained complementary relationships and dependencies between representations. Similarly, Divergence Guided Alignment proves especially beneficial when the dominant source of incompatibility lies in the distributional characteristics of the representation spaces. An equally important observation emerging from this thesis is that these alignment paradigms should not be viewed as competing alternatives. Rather, they address different forms of representational incompatibility and can often complement one another. This insight motivated the development of **PARROT**, where Geometry Guided Alignment and Interaction Guided Alignment were unified within a single framework. The success of **PARROT** demonstrates that hybrid alignment strategies can be particularly advantageous when multiple forms of heterogeneity coexist simultaneously within the representations. Consequently, a central takeaway of this thesis is that the selection of an alignment strategy should be guided by the characteristics of the representations and the requirements of the downstream task, rather than by a one-size-fits-all fusion methodology.

8.2 Future Work

Based on the findings presented in this thesis, several avenues for future research remain open. These are outlined below:

- **Multi-Encoder LALMs:** LALMs have shown sufficient development across different speech processing tasks. However, most current LALMs rely on a single audio encoder, which may limit their ability to capture important acoustic and paralinguistic cues. As demonstrated throughout this thesis, fusion of multiple PTMs consistently outperforms individual PTMs across affective and secure VIS tasks. A natural and promising extension of this work is therefore to integrate the alignment and fusion strategies proposed in this thesis into LALMs for building multi-encoder LALMs. Such integration could enable more robust emotion reasoning and more interactive, human-centered conversational systems.
- **Open-Set Models for Secure VIS:** `MiO` has demonstrated strong robustness for SSD under cross-corpus settings involving different languages, data distributions, and speech generators, suggesting encouraging generalization capabilities. However, a key challenge remains for SSSA, where speech generated by previously unseen generators may appear after deployment. Consequently, future research should move beyond the conventional closed-set formulation of SSSA and investigate adaptive and open-set settings. One promising direction is to reformulate SSSA as a source verification or source parsing problem. In particular, source parsing would extend the objective beyond identifying known generators by predicting the parameters of the unseen generators, thereby enabling their attribution and facilitating forensic tracing even when the generator was not observed during training.
- **Towards Efficient PTM Fusion-based VIS:** While fusion of heterogeneous PTMs representations consistently improves performance across affective and secure VIS tasks, deploying multiple PTMs simultaneously may increase computational and deployment costs. Therefore, a promising future direction is to leverage multi-encoder fusion systems as teacher models and distill their knowledge into lightweight single-encoder models. Such an approach could retain the performance benefits obtained through PTM complementarity while enabling more efficient and practical deployment of VIS systems.

8.3 Publications

The papers carried out as part of this thesis are outlined as follows (The papers are ordered in year-wise manner):

- *Heterogeneity over Homogeneity: Investigating Multilingual Speech Pre-Trained Models for Detecting Audio Deepfake*, Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics **(NAACL)** Findings, 2024.

- *Investigating Prosodic Signatures via Speech Pre-Trained Models for Audio Deepfake Source Attribution*, Proc. 63rd Annual Meeting of the Association for Computational Linguistics **(ACL)** Findings, 2025.
- *Strong Alone, Stronger Together: Synergizing Modality-Binding Foundation Models with Optimal Transport for Non-Verbal Emotion Recognition*, Proc. 50th IEEE International Conference on Acoustics, Speech and Signal Processing **(ICASSP)**, 2025.
- *HYFuse: Aligning Heterogeneous Speech Pre-Trained Representations in Hyperbolic Space for Speech Emotion Recognition*, Proc. 26th Annual Conference of the International Speech Communication Association **(INTERSPEECH)**, 2025.
- *PARROT: Synergizing Mamba and Attention-based SSL Pre-Trained Models via Parallel Branch Hadamard Optimal Transport for Speech Emotion Recognition*, Proc. 26th Annual Conference of the International Speech Communication Association **((INTERSPEECH)**, 2025.
- *Towards Source Attribution of Singing Voice Deepfake with Multimodal Foundation Models*, Proc. 26th Annual Conference of the International Speech Communication Association **(INTERSPEECH)**, 2025.
- *Are Multimodal Foundation Models All That Is Needed for Emofake Detection?*, Accepted to 17th Asia Pacific Signal and Information Processing Association Annual Summit and Conference **(APSIPA ASC)**, 2025.
- *CodecFake-Kids: Towards Detecting Neural Codec Synthesized Children Voice Deepfakes*, To be submitted to Transactions on Machine Learning Research (**TMLR**).

Other Works: Additionally, I have also included the titles of additional papers carried out during my PhD duration. The papers are as follows (The papers are ordered in year-wise manner):

- *Transforming the Embeddings: A Lightweight Technique for Speech Emotion Recognition Tasks*, Proc. 24th Annual Conference of the International Speech Communication Association **(INTERSPEECH)**, 2023.
- *Are Paralinguistic Representations all that is needed for Speech Emotion Recognition?*, Proc. 25th Annual Conference of the International Speech Communication Association **(INTERSPEECH)**, 2024.
- *Towards Multilingual Audio-Visual Question Answering*, Proc. 25th Annual Conference of the International Speech Communication Association **(INTERSPEECH)**, 2024.
- *SONIC: Synergizing Vision Foundation Models for Stress Recognition from ECG Signals*, Proc. 32nd European Signal Processing Conference **(EUSIPCO)**, 2024.
- *Whispers of Trauma: Leveraging Social Media for Assessing Mental Health in Victims of Childhood Sexual Abuse*, Proc. 16th International Conference on Advances in Social Networks Analysis and Mining **(ASONAM)**, 2024.

- *Towards Fusion of Neural Audio Codec-based Representations with Spectral for Heart Murmur Classification via Bandit-based Cross-Attention Mechanism*, Proc. 26th Annual Conference of the International Speech Communication Association **(INTERSPEECH)**, 2025.
- *SNIFR: Boosting Fine-Grained Child Harmful Content Detection Through Audio-Visual Alignment with Cascaded Cross-Transformer*, Proc. 26th Annual Conference of the International Speech Communication Association **(INTERSPEECH)**, 2025.
- *Investigating the Reasonable Effectiveness of Speaker Pre-Trained Models and Their Synergistic Power for SingMOS Prediction*, Proc. 26th Annual Conference of the International Speech Communication Association **(INTERSPEECH)**, 2025.
- *Towards Machine Unlearning for Paralinguistic Speech Processing*, Proc. 26th Annual Conference of the International Speech Communication Association **(INTERSPEECH)**, 2025.
- *Are Mamba-based Audio Foundation Models the Best Fit for Non-Verbal Emotion Recognition?*, Proc. 33rd European Signal Processing Conference **(EUSIPCO)**, 2025.
- *Source Tracing of Synthetic Speech Systems Through Paralinguistic Pre-Trained Representations*, Proc. 33rd European Signal Processing Conference **(EUSIPCO)**, 2025.
- *Rethinking Cross-Corpus Speech Emotion Recognition Benchmarking: Are Paralinguistic Pre-Trained Representations Sufficient?*, Proc. 17th Asia Pacific Signal and Information Processing Association Annual Summit and Conference **(APSIPA ASC)**, 2025.
- *Beyond Speech and More: Investigating the Emergent Ability of Speech Pre-Trained Models for Classifying Physiological Time-Series Signals*, Proc. 17th Asia Pacific Signal and Information Processing Association Annual Summit and Conference **(APSIPA ASC)**, 2025.
- *Investigating Polyglot Speech Foundation Models for Learning Collective Emotion from Crowds*, Proc. 17th Asia Pacific Signal and Information Processing Association Annual Summit and Conference **(APSIPA ASC)**, 2025.

REFERENCES

1. **Agarla, M.**, **S. Bianco**, **L. Celona**, **P. Napoletano**, **A. Petrovsky**, **F. Piccoli**, **R. Schettini**, and **I. Shanin** (2024). Semi-supervised cross-lingual speech emotion recognition. *Expert Systems with Applications*, **237**, 121368.

2. **Ahn, Y.**, **S. J. Lee**, and **J. W. Shin** (2021). Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation. *IEEE Signal Processing Letters*, **28**, 1190–1194.

3. **Akhtar, M. M.**, **O. C. Phukan**, **S. R. Behera**, **P. B. Reddy**, **A. C. Nayak**, **S. K. Nayak**, **A. B. Buduru**, *et al.* (2025). Are multimodal foundation models all that is needed for emofake detection? *arXiv preprint arXiv:2509.16193*.

4. **Altalahin, I.**, **S. AlZu'bi**, **A. Alqudah**, *et al.*, Unmasking the truth: A deep learning approach to detecting deepfake audio through mfcc features. *In Proc. of 2023 International Conference on Information Technology (ICIT)*. 2023.

5. **Alzantot, M.**, **Z. Wang**, and **M. B. Srivastava**, Deep Residual Neural Networks for Audio Spoofing Detection. *In Proc. Interspeech 2019*. 2019.

6. **Aouf, A.** (2019). Basic arabic vocal emotions dataset (baved). https://github.com/40uf411/Basic-Arabic-Vocal-Emotions-Dataset. GitHub repository.

7. **Araño, K. A.**, **P. Gloor**, **C. Orsenigo**, and **C. Vercellis** (2021*a*). When old meets new: emotion recognition from speech signals. *Cognitive Computation*, **13**, 771–783.

8. **Araño, K. A.**, **C. Orsenigo**, **M. Soto**, and **C. Vercellis** (2021*b*). Multimodal sentiment and emotion recognition in hyperbolic space. *Expert Systems with Applications*, **184**, 115507.

9. **Arunkumar, A.**, **V. Nileshkumar Sukhadia**, and **S. Umesh**, Investigation of Ensemble features of Self-Supervised Pretrained Models for Automatic Speech Recognition. *In Proc. Interspeech 2022*. 2022.

10. **Atmaja, B. T.** and **A. Sasou** (2022). Evaluating self-supervised speech representations for speech emotion recognition. *IEEE Access*, **10**, 124396–124407.

11. **Ba, Z.**, **Q. Wen**, **P. Cheng**, **Y. Wang**, **F. Lin**, **L. Lu**, and **Z. Liu**, Transferring audio deepfake detection capability across languages. *In Proceedings of the ACM Web Conference 2023*. 2023.

12. **Babu, A.**, **C. Wang**, **A. Tjandra**, **K. Lakhotia**, **Q. Xu**, **N. Goyal**, **K. Singh**, **P. von Platen**, **Y. Saraf**, **J. Pino**, **A. Baevski**, **A. Conneau**, and **M. Auli**, XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *In Proc. Interspeech 2022*. 2022.

13. **Baevski, A.**, **Y. Zhou**, **A. Mohamed**, and **M. Auli** (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, **33**, 12449–12460.

14. **Balamurali, B. T.**, **K. E. Lin**, **S. Lui**, **J.-M. Chen**, and **D. Herremans** (2019). Toward robust audio spoofing detection: A detailed comparison of traditional and learned features. *IEEE Access*, **7**, 84229–84241.

15. **Bhagtani, K.**, **A. K. S. Yadav**, **P. Bestagini**, and **E. J. Delp**, Attribution of diffusion based deepfake speech generators. *In 2024 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2024.

16. **Borsos, Z.**, **R. Marinier**, **D. Vincent**, **E. Kharitonov**, **O. Pietquin**, **M. Sharifi**, **D. Roblek**, **O. Teboul**, **D. Grangier**, **M. Tagliasacchi**, *et al.* (2023). Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, **31**, 2523–2533.

17. **Burkhardt, F.**, **A. Paeschke**, **M. Rolfes**, **W. F. Sendlmeier**, **B. Weiss**, *et al.*, A database of german emotional speech. *In Interspeech*, volume 5. 2005.

18. **Cahyawijaya, S.**, **H. Lovenia**, **W. Chung**, **R. Frieske**, **Z. Liu**, and **P. Fung**, Cross-Lingual Cross-Age Adaptation for Low-Resource Elderly Speech Emotion Recognition. *In Proc. INTERSPEECH 2023*. 2023.

19. **Cao, H.**, **D. G. Cooper**, **M. K. Keutmann**, **R. C. Gur**, **A. Nenkova**, and **R. Verma** (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, **5**(4), 377–390.

20. **Chen, L.-W.** and **A. I. Rudnicky** (2021). Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

21. **Chen, S.**, **C. Wang**, **Z. Chen**, **Y. Wu**, **S. Liu**, **Z. Chen**, **J. Li**, **N. Kanda**, **T. Yoshioka**, **X. Xiao**, *et al.* (2022*a*). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1505–1518.

22. **Chen, S.**, **Y. Wu**, **C. Wang**, **Z. Chen**, **Z. Chen**, **S. Liu**, **J. Wu**, **Y. Qian**, **F. Wei**, **J. Li**, *et al.*, Unispeech-sat: Universal speech representation learning with speaker aware pre-training. *In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022*b*.

23. **Chen, X.**, **J. Du**, **H. Wu**, **L. Zhang**, **I. Lin**, **I. Chiu**, **W. Ren**, **Y. Tseng**, **Y. Tsao**, **J.-S. R. Jang**, *et al.* (2025*a*). Codecfake+: A large-scale neural audio codec-based deepfake speech dataset. *arXiv preprint arXiv:2501.08238*.

24. **Chen, X.**, **I. Lin**, **L. Zhang**, **H. Wu**, **H.-y. Lee**, **J.-S. R. Jang**, *et al.* (2025*b*). Towards generalized source tracing for codec-based deepfake speech. *arXiv preprint arXiv:2506.07294*.

25. **Chen, X.**, **H. Wu**, **R. Jang**, and **H. yi Lee**, Singing voice graph modeling for singfake detection. *In Interspeech 2024*. 2024. ISSN 2958-1796.

26. **Chetia Phukan, O.**, **A. Balaji Buduru**, and **R. Sharma**, Transforming the Embeddings: A Lightweight Technique for Speech Emotion Recognition Tasks. *In Proc. INTERSPEECH 2023*. 2023.

27. **Chetia Phukan, O.**, **D. Singh**, **S. R. Behera**, **A. B. Buduru**, and **R. Sharma**, Investigating prosodic signatures via speech pre-trained models for audio deepfake source attribution. *In* **W. Che**, **J. Nabende**, **E. Shutova**, and **M. T. Pilehvar** (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Vienna, Austria, 2025. ISBN 979-8-89176-256-5. URL `https://aclanthology.org/2025.findings-acl.218/`.

28. **Chhibber, M.**, **J. Mishra**, **H.-J. Shim**, and **T. H. Kinnunen**, An explainable probabilistic attribute embedding approach for spoofed speech characterization. *In ICASSP*

2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2025.

29. **Chung, J.**, **T. Zhu**, **M. G. Saez-Diez**, **J. C. Niebles**, **H. Zhou**, and **O. Russakovsky**, Unifying specialized visual encoders for video language models. *In Forty-second International Conference on Machine Learning*. 2025. URL `https://openreview.net/forum?id=oYzDcCZdR9`.

30. **Conneau, A.**, **A. Baevski**, **R. Collobert**, **A. Mohamed**, and **M. Auli**, Unsupervised Cross-Lingual Representation Learning for Speech Recognition. *In Proc. Interspeech 2021*. 2021.

31. **Conti, E.**, **D. Salvi**, **C. Borrelli**, **B. Hosler**, **P. Bestagini**, **F. Antonacci**, **A. Sarti**, **M. C. Stamm**, and **S. Tubaro**, Deepfake speech detection through emotion recognition: a semantic approach. *In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

32. **Cowie, R.**, **E. Douglas-Cowie**, **N. Tsapatsoulis**, **G. Votsis**, **S. Kollias**, **W. Fellenz**, and **J. G. Taylor** (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, **18**(1), 32–80.

33. **Défossez, A.**, **J. Copet**, **G. Synnaeve**, and **Y. Adi** (2022). High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

34. **Défossez, A.**, **J. Copet**, **G. Synnaeve**, and **Y. Adi** (2023). High fidelity neural audio compression. *Transactions on Machine Learning Research*. ISSN 2835-8856. URL `https://openreview.net/forum?id=ivCd8z8zR2`. Featured Certification, Reproducibility Certification.

35. **Diatlova, D.**, **A. Udalov**, **V. Shutov**, and **E. Spirin** (2024). Adapting wavlm for speech emotion recognition. *ArXiv*, **abs/2405.04485**.

36. **Don-Yehiya, S.**, **E. Venezian**, **C. Raffel**, **N. Slonim**, and **L. Choshen**, Cold fusion: Collaborative descent for distributed multitask finetuning. *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023.

37. **Du, Z.**, **S. Zhang**, **K. Hu**, and **S. Zheng**, Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. *In ICASSP 2024-2024 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.

38. **Duville, M. M.**, **L. M. Alonso-Valerdi**, and **D. I. Ibarra-Zarate**, The mexican emotional speech database (mesd): elaboration and assessment based on machine learning. *In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021.

39. **Egas-López, J. V.**, **G. Kiss**, **D. Sztahó**, and **G. Gosztolya**, Automatic assessment of the degree of clinical depression from speech using x-vectors. *In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022.

40. **Elizalde, B.**, **S. Deshmukh**, **M. Al Ismail**, and **H. Wang**, Clap learning audio concepts from natural language supervision. *In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

41. **Elizalde, B.**, **S. Deshmukh**, and **H. Wang**, Natural language supervision for general-purpose audio representations. *In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

42. **Eom, Y.**, **Y. Lee**, **J. S. Um**, and **H. Kim** (2022). Anti-spoofing using transfer learning with variational information bottleneck. *arXiv preprint arXiv:2204.01387*.

43. **Erol, M. H.**, **A. Senocak**, **J. Feng**, and **J. S. Chung** (2024). Audio mamba: Bidirectional state space model for audio representation learning. *IEEE Signal Processing Letters*, **31**, 2975–2979.

44. **Fukumori, T.**, **T. Ishida**, and **Y. Yamashita**, Investigating the effectiveness of speaker embeddings for shout intensity prediction. *In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023.

45. **Gao, J.**, **P. Li**, **Z. Chen**, and **J. Zhang** (2020). A survey on deep learning for multimodal data fusion. *Neural computation*, **32**(5), 829–864.

46. **Girdhar, R.**, **A. El-Nouby**, **Z. Liu**, **M. Singh**, **K. V. Alwala**, **A. Joulin**, and **I. Misra**, Imagebind: One embedding space to bind them all. *In CVPR*. 2023.

47. **Gomez-Alanis, A.**, **A. M. Peinado**, **J. A. Gonzalez**, and **A. M. Gomez**, A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection. *In Proc. Interspeech 2019*. 2019.

48. **Gu, A.** and **T. Dao** (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

49. **Guo, Y.**, **H. Huang**, **X. Chen**, **H. Zhao**, and **Y. Wang**, Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier. *In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

50. **Guragain, A.**, **T. Liu**, **Z. Pan**, **H. B. Sailor**, and **Q. Wang**, Speech foundation model ensembles for the controlled singing voice deepfake detection (ctrsvdd) challenge 2024. *In 2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024.

51. **Hamza, A.**, **A. R. R. Javed**, **F. Iqbal**, *et al.* (2022). Deepfake audio detection via mfcc features using machine learning. *IEEE Access*, **10**, 134018–134028.

52. **Heracleous, P.**, **S. Fukayama**, **J. Ogata**, and **Y. Mohammad**, Applying generative adversarial networks and vision transformers in speech emotion recognition. *In HCI International 2022-Late Breaking Papers. Multimodality in Advanced Interaction Environments: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings*. Springer, 2022.

53. **Holz, N.**, **P. Larrouy-Maestri**, and **D. Poeppel** (2022). The variably intense vocalizations of affect and emotion (vivae) corpus prompts new perspective on nonspeech perception. *Emotion*, **22**(1), 213.

54. **Hsu, J.-H.**, **M.-H. Su**, **C.-H. Wu**, and **Y.-H. Chen** (2021*a*). Speech emotion recognition considering nonverbal vocalization in affective conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 1675–1686.

55. **Hsu, W.-N.**, **B. Bolte**, **Y.-H. H. Tsai**, **K. Lakhotia**, **R. Salakhutdinov**, and **A. Mohamed** (2021*b*). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 3451–3460.

56. **Huang, K.-P.**, **T.-H. Feng**, **Y.-K. Fu**, **T.-Y. Hsu**, **P.-C. Yen**, **W.-C. Tseng**, **K.-W. Chang**, and **H.-Y. Lee**, Ensemble knowledge distillation of self-supervised speech models. *In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

57. **Huang, Z.**, **M. Dong**, **Q. Mao**, and **Y. Zhan**, Speech emotion recognition using cnn. *In Proceedings of the 22nd ACM international conference on Multimedia*. 2014.

58. **Ilias, L.** and **D. Askounis**, A Cross-Attention Layer coupled with Multimodal Fusion Methods for Recognizing Depression from Spontaneous Speech. *In Interspeech 2024*. 2024. ISSN 2958-1796.

59. **Issa, D.**, **M. F. Demirci**, and **A. Yazici** (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, **59**, 101894.

60. **Jiang, Z.**, **H. Zhu**, **L. Peng**, **W. Ding**, and **Y. Ren**, Self-Supervised Spoofing Audio Detection Scheme. *In Proc. Interspeech 2020*. 2020.

61. **Jin, Q.**, **C. Li**, **S. Chen**, and **H. Wu**, Speech emotion recognition with acoustic and lexical features. *In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015.

62. **jin Shim, H.**, **H.-S. Heo**, **J. weon Jung**, and **H.-J. Yu**, Self-Supervised Pre-Training with Acoustic Configurations for Replay Spoofing Detection. *In Proc. Interspeech 2020*. 2020.

63. **Jing, X.**, **A. Triantafyllopoulos**, and **B. Schuller**, ParaCLAP – Towards a general language-audio model for computational paralinguistic tasks. *In Interspeech 2024*. 2024. ISSN 2958-1796.

64. **Jung, J.-w.**, **H.-S. Heo**, **H. Tak**, **H.-j. Shim**, **J. S. Chung**, **B.-J. Lee**, **H.-J. Yu**, and **N. Evans**, Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. *In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

65. **Kawa, P.**, **M. Plata**, **M. Czuba**, **P. Szymański**, and **P. Syga**, Improved DeepFake Detection Using Whisper Features. *In Proc. INTERSPEECH 2023*. 2023.

66. **Keesing, A.**, **Y. S. Koh**, and **M. Witbrock**, Acoustic features and neural representations for categorical emotion recognition from speech. *In Interspeech*. 2021.

67. **Kesarwani, A.**, **S. Das**, **D. R. Kisku**, and **M. Dalui** (2024). Dual mode information fusion with pre-trained cnn models and transformer for video-based non-invasive anaemia detection. *Biomedical Signal Processing and Control*, **88**, 105592.

68. **Khan, M.**, **W. Gueaieb**, **A. El Saddik**, and **S. Kwon** (2024). Mser: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Systems with Applications*, **245**, 122946.

69. **Kilinc, H. H.** and **F. Kaledibi**, Audio deepfake detection by using machine and deep learning. *In Proc. of 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*. 2023.

70. **Kinnunen, T.**, **M. Sahidullah**, **H. Delgado**, **N. Evans**, **M. Todisco**, *et al.*, The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. *In Proc. of INTERSPEECH*. 2017.

71. **Kishore, K. K.** and **P. K. Satish**, Emotion recognition in speech using mfcc and wavelet features. *In 2013 3rd IEEE International Advance Computing Conference (IACC)*. IEEE, 2013.

72. **Klein, N.**, **T. Chen**, **H. Tak**, **R. Casal**, and **E. Khoury**, Source tracing of audio deepfake systems. *In Interspeech 2024*. 2024. ISSN 2958-1796.

73. **Kornblith, S.**, **M. Norouzi**, **H. Lee**, and **G. Hinton**, Similarity of neural network representations revisited. *In International conference on machine learning*. PMlR, 2019.

74. **Kulkarni, A.**, **F. Teixeira**, **E. Hermann**, **T. Rolland**, **I. Trancoso**, and **M. M. Doss** (2025). Children's voice privacy: First steps and emerging challenges. *arXiv preprint arXiv:2506.00100*.

75. **Kumar, G. K.** and **K. Nandakumar**, Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *In Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*. 2022.

76. **Kumar, R.**, **P. Seetharaman**, **A. Luebs**, **I. Kumar**, and **K. Kumar** (2024). High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, **36**.

77. **Landry, D.**, **Q. He**, **H. Yan**, and **Y. Li** (2020). Asvp-esd: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances. *Global Scientific Journals*, **8**, 1793–1798.

78. **Latif, S.**, **A. Qayyum**, **M. Usman**, and **J. Qadir**, Cross lingual speech emotion recognition: Urdu vs. western languages. *In 2018 International conference on frontiers of information technology (FIT)*. IEEE, 2018.

79. **Lavrentyeva, G.**, **S. Novoselov**, **A. Tseren**, **M. Volkova**, **A. Gorlanov**, and **A. Kozlov** (2019). Stc antispoofing systems for the asvspoof2019 challenge. *arXiv preprint arXiv:1904.05576*.

80. **Lee, C. M.** and **S. S. Narayanan** (2005). Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, **13**(2), 293–303.

81. **Lee, Y.**, **N. Kim**, **J. Jeong**, and **I.-Y. Kwak** (2023). Experimental case study of self-supervised learning for voice spoofing detection. *IEEE Access*, **11**, 24216–24226.

82. **Lei, J.**, **X. Zhu**, and **Y. Wang** (2022). Bat: Block and token self-attention for speech emotion recognition. *Neural Networks*, **156**, 67–80.

83. **Li, Y.**, **R. Yuan**, **G. Zhang**, **Y. Ma**, **X. Chen**, **H. Yin**, **C. Xiao**, **C. Lin**, **A. Ragni**, **E. Benetos**, *et al.* (2023). Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*.

84. **Li, Y.**, **R. Yuan**, **G. Zhang**, **Y. Ma**, **C. Lin**, **X. Chen**, **A. Ragni**, **H. Yin**, **Z. Hu**, **H. He**, **E. Benetos**, **N. Gyenge**, **R. Liu**, and **J. Fu** (2022). Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning. *ArXiv*, **abs/2212.02508**.

85. **Lian, Z.**, **B. Liu**, and **J. Tao** (2021). Ctnet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 985–1000.

86. **Lin, W.-C.**, **S. Ghaffarzadegan**, **L. Bondi**, **A. Kumar**, **S. Das**, and **H.-H. Wu**, Clap4emo: Chatgpt-assisted speech emotion retrieval with natural language supervision. *In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

87. **Liu, X.**, **X. Wang**, **M. Sahidullah**, *et al.* (2023). ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

88. **Lu, Y.**, **Y. Xie**, **R. Fu**, **Z. Wen**, **J. Tao**, **Z. Wang**, **X. Qi**, **X. Liu**, **Y. Li**, **Y. Liu**, **X. Wang**, and **S. Shi**, Codecfake: An initial dataset for detecting llm-based deepfake audio. *In Interspeech 2024*. 2024. ISSN 2958-1796.

89. **Luo, A.**, **E. Li**, **Y. Liu**, **X. Kang**, and **Z. J. Wang**, A capsule network based approach for detection of audio spoofing attacks. *In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

90. **Ma, H.**, **J. Yi**, **C. Wang**, **X. Yan**, **J. Tao**, **T. Wang**, **S. Wang**, and **R. Fu** (2024). Cfad: A chinese dataset for fake audio detection. *Speech Communication*, **164**, 103122.

91. **Ma, X.**, **T. Liang**, **S. Zhang**, **S. Huang**, and **L. He**, Improved lightcnn with attention modules for asv spoofing detection. *In 2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021.

92. **Ma, X.**, **S. Zhang**, **S. Huang**, **J. Gao**, **Y. Hu**, and **L. He**, How to boost anti-spoofing with x-vectors. *In 2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023.

93. **Mao, S.**, **D. Tao**, **G. Zhang**, **P. Ching**, and **T. Lee**, Revisiting hidden markov models for speech emotion recognition. *In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

94. **Martín-Doñas, J. M.** and **A. Álvarez**, The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. *In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

95. **Mena, C. D. H.**, **D. E. Mollberg**, **M. Borskỳ**, and **J. Guðnason**, Samrómur children: An icelandic speech corpus. *In Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2022.

96. **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. *IEEE Journal of Selected Topics in Signal Processing*.

97. **Morais, E.**, **R. Hoory**, **W. Zhu**, **I. Gat**, **M. Damasceno**, and **H. Aronowitz**, Speech emotion recognition using self-supervised features. *In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

98. **Mousavi, P.**, **L. Della Libera**, **J. Duret**, **A. Ploujnikov**, **C. Subakan**, and **M. Ravanelli** (2024). Dasb–discrete audio and speech benchmark. *arXiv preprint arXiv:2406.14294*.

99. **Müller, N.**, **P. Czempin**, **F. Diekmann**, **A. Froghyar**, and **K. Böttinger**, Does Audio Deepfake Detection Generalize? *In Proc. Interspeech 2022*. 2022*a*.

100. **Müller, N.**, **F. Diekmann**, and **J. Williams**, Attacker attribution of audio deepfakes. *In Interspeech 2022*. 2022*b*. ISSN 2958-1796.

101. **Nogueiras, A.**, **A. Moreno**, **A. Bonafonte**, and **J. B. Mariño**, Speech emotion recognition using hidden markov models. *In Seventh European conference on speech communication and technology*. 2001.

102. **Osman, M.**, **D. Z. Kaplan**, and **T. Nadeem**, Ser evals: In-domain and out-of-domain benchmarking for speech emotion recognition. *In Interspeech 2024*. 2024. ISSN 2958-1796.

103. **Ottl, S.**, **S. Amiriparian**, **M. Gerczuk**, **V. Karas**, and **B. Schuller**, Group-level speech emotion recognition utilising deep spectrum features. *In Proceedings of the 2020 International Conference on Multimodal Interaction*. 2020.

104. **Pan, Y.**, **Y. Hu**, **Y. Yang**, **W. Fei**, **J. Yao**, **H. Lu**, **L. Ma**, and **J. Zhao**, Gemo-clap: Gender-attribute-enhanced contrastive language-audio pretraining for accurate speech emotion recognition. *In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

105. **Parry, J.**, **D. Palaz**, **G. Clarke**, **P. Lecomte**, **R. Mead**, **M. Berger**, and **G. Hofer**, Analysis of deep learning architectures for cross-corpus speech emotion recognition. *In Interspeech*. 2019.

106. **Pasad, A.** (2025). What do speech foundation models learn? analysis and applications. *arXiv preprint arXiv:2508.12255*.

107. **Pascu, O.**, **A. Stan**, **D. Oneata**, **E. Oneata**, and **H. Cucu**, Towards generalisable and calibrated audio deepfake detection with self-supervised representations. *In Interspeech 2024*. 2024. ISSN 2958-1796.

108. **Peng, J.** and **G. Seetharaman**, Chernoff distance and relief feature selection. *In 2012 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2012.

109. **Pepino, L.**, **P. Riera**, and **L. Ferrer**, Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. *In Proc. Interspeech 2021*. 2021.

110. **Phukan, O. C.**, **M. M. Akhtar**, **S. R. Behera**, **S. Kalita**, **A. B. Buduru**, **R. Sharma**, **S. M. Prasanna**, *et al.*, Strong alone, stronger together: Synergizing modality-binding foundation models with optimal transport for non-verbal emotion recognition. *In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025*a*.

111. **Phukan, O. C.**, **M. M. Akhtar**, **Girish**, **S. R. Behera**, **J. S. K. Patibandla**, **A. B. Buduru**, and **R. Sharma**, PARROT: Synergizing Mamba and Attention-based SSL Pre-Trained Models via Parallel Branch Hadamard Optimal Transport for Speech Emotion Recognition. *In Interspeech 2025*. 2025*b*. ISSN 2958-1796.

112. **Phukan, O. C.**, **Girish**, **M. M. Akhtar**, **S. R. Behera**, **P. Mallick**, **P. B. Reddy**, **A. B. Buduru**, and **R. Sharma**, Towards Source Attribution of Singing Voice Deepfake with Multimodal Foundation Models. *In Interspeech 2025*. 2025*c*. ISSN 2958-1796.

113. **Phukan, O. C.**, **Girish**, **M. M. Akhtar**, **S. R. Behera**, **P. B. Reddy**, **A. B. Buduru**, and **R. Sharma**, HYFuse: Aligning Heterogeneous Speech Pre-Trained Representations in Hyperbolic Space for Speech Emotion Recognition. *In Interspeech 2025*. 2025*d*. ISSN 2958-1796.

114. **Phukan, O. C.**, **G. S. Kashyap**, **A. B. Buduru**, and **R. Sharma**, Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake. *In NAACL-HLT*. 2024*a*.

115. **Phukan, O. C.**, **G. S. Kashyap**, **A. B. Buduru**, and **R. Sharma** (2024*b*). Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake. *arXiv preprint arXiv:2404.00809*.

116. **Phukan, O. C.**, **M. Mujtaba Akhtar**, **Girish**, **S. Ranjan Behera**, **S. Kalita**, **A. B. Buduru**, **R. Sharma**, and **S. M. Prasanna**, Strong alone, stronger together: Synergizing modality-binding foundation models with optimal transport for non-verbal emotion recognition. *In ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025*e*.

117. **Phukan, O. C.**, **D. Singh**, **S. R. Behera**, **A. B. Buduru**, and **R. Sharma**, Investigating prosodic signatures via speech pre-trained models for audio deepfake source attribution. *In Findings of the Association for Computational Linguistics: ACL 2025*. 2025*f*.

118. **Pramanick, S.**, **A. B. Roy**, and **V. M. Patel** (2021). Multimodal learning using optimal transport for sarcasm and humor detection. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 546–556.

119. **Pratap, V.**, **A. Tjandra**, **B. Shi**, **P. Tomasello**, **A. Babu**, **S. Kundu**, **A. Elkahky**, **Z. Ni**, **A. Vyas**, **M. Fazel-Zarandi**, *et al.* (2023). Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

120. **Praveen, R. G.** and **J. Alam**, Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

121. **Qian, Y.**, **N. Chen**, and **K. Yu** (2016). Deep features for automatic spoofing detection. *Speech Communication*, **85**, 43–52.

122. **Radford, A.**, **J. W. Kim**, **T. Xu**, **G. Brockman**, **C. McLeavey**, and **I. Sutskever**, Robust speech recognition via large-scale weak supervision. *In International Conference on Machine Learning*. PMLR, 2023.

123. **Radha, K.**, **M. Bansal**, and **R. B. Pachori** (2024). Automatic speaker and age identification of children from raw speech using sincnet over erb scale. *Speech Communication*, **159**, 103069.

124. **Ranjan, R.**, **M. Vatsa**, and **R. Singh**, Statnet: Spectral and temporal features based multi-task network for audio spoofing detection. *In 2022 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2022.

125. **Ravanelli, M.**, **T. Parcollet**, **P. Plantinga**, **A. Rouhe**, **S. Cornell**, **L. Lugosch**, **C. Subakan**, **N. Dawalatabad**, **A. Heba**, **J. Zhong**, **J.-C. Chou**, **S.-L. Yeh**, **S.-W. Fu**, **C.-F.**

Liao, **E. Rastorgueva**, **F. Grondin**, **W. Aris**, **H. Na**, **Y. Gao**, **R. D. Mori**, and **Y. Bengio** (2021). SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

126. **Ren, W.**, **Y.-C. Lin**, **H.-C. Chou**, **H. Wu**, **Y.-C. Wu**, **C.-C. Lee**, **H.-y. Lee**, and **Y. Tsao** (2024). Emo-codec: An in-depth look at emotion preservation capacity of legacy and neural codec models with subjective and objective evaluations. *arXiv preprint arXiv:2407.15458*.

127. **Sahidullah, M.**, **T. Kinnunen**, and **C. Hanilçi**, A comparison of features for synthetic speech detection. *In Proc. Interspeech 2015*. 2015.

128. **Schuller, B.**, **G. Rigoll**, and **M. Lang**, Hidden markov model-based speech emotion recognition. *In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2. Ieee, 2003.

129. **Shaaban, O. A.**, **R. Yildirim**, and **A. A. Alguttar** (2023). Audio deepfake approaches. *IEEE Access*, **11**, 132652–132682.

130. **Shams, S.**, **S. S. Dindar**, **X. Jiang**, and **N. Mesgarani** (2024). Ssamba: Self-supervised audio representation learning with mamba state space model. *arXiv preprint arXiv:2405.11831*.

131. **Shankar, S.**, Multimodal fusion via cortical network inspired losses. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022.

132. **Sharma, G.**, **R. Chinmay**, and **R. Sharma**, Late fusion of transformers for sentiment analysis of code-switched data. *In Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023.

133. **Shi, J.**, **Y. Zhang**, and **Y. Gao**, CLEP-DG: Contrastive Learning for Speech Emotion Domain Generalization via Soft Prompt Tuning. *In Interspeech 2025*. 2025. ISSN 2958-1796.

134. **Siuzdak, H.**, **F. Grötschla**, and **L. A. Lanzendörfer**, Snac: Multi-scale neural audio codec. *In Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*. 2024.

135. **Snyder, D.**, **D. Garcia-Romero**, **G. Sell**, **D. Povey**, and **S. Khudanpur**, X-vectors: Robust dnn embeddings for speaker recognition. *In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018*a*.

136. **Snyder, D.**, **D. Garcia-Romero**, **G. Sell**, **D. Povey**, and **S. Khudanpur**, X-vectors: Robust dnn embeddings for speaker recognition. *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018*b*.

137. **Sobti, R.**, **K. Guleria**, and **V. Kadyan** (2024). Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges. *Multimedia Tools and Applications*, **83**(35), 81933–81995.

138. **Steyaert, S.**, **M. Pizurica**, **D. Nagaraj**, **P. Khandelwal**, **T. Hernandez-Boussard**, **A. J. Gentles**, and **O. Gevaert** (2023). Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature machine intelligence*, **5**(4), 351–362.

139. **Todisco, M.**, **X. Wang**, **V. Vestman**, **M. Sahidullah**, and **K. Lee**, ASVspoof 2019: Future horizons in spoofed and fake audio detection. *In Proc. of INTERSPEECH*. 2019.

140. **Tom, F.**, **M. Jain**, and **P. Dey**, End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention. *In Proc. Interspeech 2018*. 2018.

141. **Tzirakis, P.**, **A. Baird**, **J. A. Brooks**, **C. Gagne**, **L. Kim**, **M. Opara**, **C. B. Gregory**, **J. Metrick**, **G. Boseck**, **V. R. Tiruvadi**, **B. Schuller**, **D. Keltner**, and **A. S. Cowen** (2023). Large-scale nonverbal vocalization detection using transformers. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

142. **Van Erven, T.** and **P. Harremos** (2014). Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, **60**(7), 3797–3820.

143. **Villalba, J.**, **A. Miguel**, **A. Ortega**, and **E. Lleida**, Spoofing detection with dnn and one-class svm for the asvspoof 2015 challenge. *In Proc. Interspeech*, volume 2015. 2015.

144. **Vryzas, N.**, **R. Kotsakis**, **A. Liatsou**, **C. A. Dimoulas**, and **G. Kalliris** (2018). Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society*, **66**(6), 457–467.

145. **Wan, F.**, **X. Huang**, **D. Cai**, **X. Quan**, **W. Bi**, and **S. Shi**, Knowledge fusion of large language models. *In The Twelfth International Conference on Learning Representations*.

146. **Wan, F.**, **X. Huang**, **D. Cai**, **X. Quan**, **W. Bi**, and **S. Shi**, Knowledge fusion of large language models. *In The Twelfth International Conference on Learning Representations*. 2024. URL `https://openreview.net/forum?id=jiDsk12qcz`.

147. **Wang, C.**, **J. Yi**, **J. Tao**, **C. Y. Zhang**, **S. Zhang**, and **X. Chen**, Detection of Cross-Dataset Fake Audio Based on Prosodic and Pronunciation Features. *In Proc. INTERSPEECH 2023*. 2023.

148. **Wang, X.**, **M. Wang**, **W. Qi**, **W. Su**, **X. Wang**, and **H. Zhou**, A novel end-to-end speech emotion recognition network with stacked transformer layers. *In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

149. **Wang, X.**, **J. Yamagishi**, **M. Todisco**, **H. Delgado**, **A. Nautsch**, **N. Evans**, **M. Sahidullah**, **V. Vestman**, **T. Kinnunen**, **K. A. Lee**, *et al.* (2020*a*). Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, **64**, 101114.

150. **Wang, Y.**, **W. Huang**, **F. Sun**, **T. Xu**, **Y. Rong**, and **J. Huang** (2020*b*). Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, **33**, 4835–4845.

151. **Wang, Z.**, **D. Ye**, **J. Li**, and **J. Deng**, Generalize audio deepfake algorithm recognition via attribution enhancement. *In ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2025.

152. **wen Yang, S.**, **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. tik Lee**, **D.-R. Liu**, **Z. Huang**, **S. Dong**, **S.-W. Li**, **S. Watanabe**, **A. Mohamed**, and **H. yi Lee**, SUPERB: Speech Processing Universal PERformance Benchmark. *In Proc. Interspeech 2021*. 2021.

153. **Wu, H.**, **H.-L. Chung**, **Y.-C. Lin**, **Y.-K. Wu**, **X. Chen**, **Y.-C. Pai**, **H.-H. Wang**, **K.-W. Chang**, **A. Liu**, and **H.-y. Lee**, Codec-SUPERB: An in-depth analysis of sound

codec models. *In* **L.-W. Ku**, **A. Martins**, and **V. Srikumar** (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand, 2024*a*. URL https://aclanthology.org/2024.findings-acl.616/.

154. **Wu, H.**, **Y. Tseng**, and **H. yi Lee**, Codecfake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems. *In Interspeech 2024*. 2024*b*. ISSN 2958-1796.

155. **Wu, Y.**, **P. Yue**, **C. Cheng**, and **T. Li**, Investigation of ensemble of self-supervised models for speech emotion recognition. *In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023*a*.

156. **Wu, Y.**, **P. Yue**, **C. Cheng**, and **T. Li**, Investigation of ensemble of self-supervised models for speech emotion recognition. *In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2023*b*.

157. **Wu, Y.-C.**, **I. D. Gebru**, **D. Marković**, and **A. Richard**, Audiodec: An open-source streaming high-fidelity neural audio codec. *In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023*c*.

158. **Wu, Z.**, **T. Kinnunen**, **N. Evans**, **J. Yamagishi**, **C. Hanilçi**, **M. Sahidullah**, and **A. Sizov**, Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. *In Sixteenth annual conference of the international speech communication association*. 2015*a*.

159. **Wu, Z.**, **T. Kinnunen**, **N. Evans**, **J. Yamagishi**, **C. Hanilçi**, *et al.*, ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. *In Proc. of INTERSPEECH*. 2015*b*.

160. **Xie, Y.**, **R. Fu**, **Z. Wen**, **Z. Wang**, **X. Wang**, **H. Cheng**, **L. Ye**, and **J. Tao**, Generalized source tracing: Detecting novel audio deepfake algorithm with real emphasis and fake dispersion strategy. *In Interspeech 2024*. 2024. ISSN 2958-1796.

161. **Xie, Y.**, **Y. Lu**, **R. Fu**, **Z. Wen**, **Z. Wang**, **J. Tao**, **X. Qi**, **X. Wang**, **Y. Liu**, **H. Cheng**, *et al.* (2025*a*). The codecfake dataset and countermeasures for the universally detection of deepfake audio. *IEEE Transactions on Audio, Speech and Language Processing*.

162. **Xie, Y.**, **X. Wang**, **Z. Wang**, **R. Fu**, **Z. Wen**, **S. Cao**, **L. Ma**, **C. Li**, **H. Cheng**, and **L. Ye** (2025*b*). Neural codec source tracing: Toward comprehensive attribution in open-set condition. *arXiv preprint arXiv:2501.06514*.

163. **Xin, D.**, **J. Jiang**, **S. Takamichi**, **Y. Saito**, **A. Aizawa**, and **H. Saruwatari** (2024). Jvnv: A corpus of japanese emotional speech with verbal content and nonverbal expressions. *IEEE Access*.

164. **Xin, D.**, **S. Takamichi**, and **H. Saruwatari** (2023). Jnv corpus: A corpus of japanese nonverbal vocalizations with diverse phrases and emotions. *Speech Commun.*, **156**, 103004.

165. **Xuan, X.**, **Y. Xiao**, **R. K. Das**, and **T. Kinnunen**, Multilingual Source Tracing of Speech Deepfakes: A First Benchmark. *In 5th Symposium on Security and Privacy in Speech Communication*. 2025.

166. **Yadav, S.** and **Z.-H. Tan**, Audio mamba: Selective state spaces for self-supervised audio representations. *In Interspeech 2024*. 2024. ISSN 2958-1796.

167. **Yamagishi, J.**, **X. Wang**, **M. Todisco**, **M. Sahidullah**, **J. Patino**, **A. Nautsch**, **X. Liu**, **K. A. Lee**, **T. Kinnunen**, and **N. Evans**, ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. *In Proc. of INTERSPEECH*. 2021.

168. **Yan, X.**, **J. Yi**, **J. Tao**, **C. Wang**, **H. Ma**, **Z. Tian**, and **R. Fu** (2022*a*). System fingerprints detection for deepfake audio: An initial dataset and investigation. *CoRR*.

169. **Yan, X.**, **J. Yi**, **J. Tao**, **C. Wang**, **H. Ma**, **T. Wang**, **S. Wang**, and **R. Fu**, An initial investigation for detecting vocoder fingerprints of fake audio. *In Proc. of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. 2022*b*.

170. **Yang, J.**, Ensemble deep learning with hubert for speech emotion recognition. *In 2023 IEEE 17th International Conference on Semantic Computing (ICSC)*. 2023.

171. **Yang, Y.**, **H. Qin**, **H. Zhou**, **C. Wang**, **T. Guo**, **K. Han**, and **Y. Wang**, A robust audio deepfake detection system via multi-view feature. *In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

172. **Yi, J.**, **R. Fu**, **J. Tao**, *et al.*, ADD 2022: The first audio deep synthesis detection challenge. *In Proc. of ICASSP*. 2022.

173. **Yi, J.**, **J. Tao**, **R. Fu**, *et al.* (2023). ADD 2023: The second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*.

174. **Yun, H.**, **Y. Kim**, and **K. Jung**, Modality alignment between deep representations for effective video-and-language learning. *In Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2022.

175. **Zaiem, S.**, **Y. Kemiche**, **T. Parcollet**, **S. Essid**, and **M. Ravanelli**, Speech Self-Supervised Representation Benchmarking: Are We Doing it Right? *In Proc. INTERSPEECH 2023*. 2023.

176. **Zang, Y.**, **J. Shi**, **Y. Zhang**, **R. Yamamoto**, **J. Han**, **Y. Tang**, **S. Xu**, **W. Zhao**, **J. Guo**, **T. Toda**, and **Z. Duan**, Ctrsvdd: A benchmark dataset and baseline analysis for controlled singing voice deepfake detection. *In Interspeech 2024*. 2024*a*. ISSN 2958-1796.

177. **Zang, Y.**, **Y. Zhang**, **M. Heydari**, and **Z. Duan**, Singfake: Singing voice deepfake detection. *In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024*b*.

178. **Zeghidour, N.**, **A. Luebs**, **A. Omran**, **J. Skoglund**, and **M. Tagliasacchi** (2021). Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **30**, 495–507.

179. **Zhai, J.** and **W. Zheng**, Fusion of features from multiple asr pretrained models: Enhancing robustness and accuracy. *In Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence*. 2024.

180. **Zhang, C. Y.**, **J. Yi**, **J. Tao**, **C. Wang**, and **X. Yan** (2023). Distinguishing neural speech synthesis models through fingerprints in speech waveforms. *ArXiv*, **abs/2309.06780**. URL `https://api.semanticscholar.org/CorpusID:261705832`.

181. **Zhang, C. Y.**, **J. Yi**, **J. Tao**, **C. Wang**, and **X. Yan**, Distinguishing neural speech synthesis models through fingerprints in speech waveforms. *In China National Conference on Chinese Computational Linguistics*. Springer, 2024*a*.

182. **Zhang, H.**, **R. Gou**, **J. Shang**, **F. Shen**, **Y. Wu**, and **G. Dai** (2021). Pre-trained deep convolution neural network model with attention for speech emotion recognition. *Frontiers in Physiology*, **12**, 643202.

183. **Zhang, X.**, **D. Zhang**, **S. Li**, **Y. Zhou**, and **X. Qiu**, Speechtokenizer: Unified speech tokenizer for speech language models. *In The Twelfth International Conference on Learning Representations*. 2024*b*. URL `https://openreview.net/forum?id=AF9Q8Vip84`.

184. **Zhang, Y.**, **Y. Zang**, **J. Shi**, **R. Yamamoto**, **J. Han**, **Y. Tang**, **T. Toda**, and **Z. Duan** (2024*c*). Svdd challenge 2024: A singing voice deepfake detection challenge (ctrsvdd track, training/development set). URL `https://doi.org/10.5281/zenodo.10467648`.

185. **Zhang, Y.**, **Y. Zang**, **J. Shi**, **R. Yamamoto**, **T. Toda**, and **Z. Duan**, Svdd 2024: The inaugural singing voice deepfake detection challenge. *In 2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024*d*.

186. **Zhao, F.**, **C. Zhang**, and **B. Geng** (2024*a*). Deep multimodal data fusion. *ACM computing surveys*, **56**(9), 1–36.

187. **Zhao, Y.**, **J. Yi**, **J. Tao**, **C. Wang**, and **Y. Dong**, Emofake: An initial dataset for emotion fake audio detection. *In China National Conference on Chinese Computational Linguistics*. Springer, 2024*b*.

188. **Zhou, J.**, **S. Wang**, **S. Zhao**, **J. He**, **H. Sun**, **H. Wang**, **C. Liu**, **A. Kong**, **Y. Guo**, **X. Yang**, *et al.* (2024). Childmandarin: A comprehensive mandarin speech dataset for young children aged 3-5. *arXiv preprint arXiv:2409.18584*.

189. **Zhu, B.**, **B. Lin**, **M. Ning**, **Y. Yan**, **J. Cui**, **W. HongFa**, **Y. Pang**, **W. Jiang**, **J. Zhang**, **Z. Li**, **C. W. Zhang**, **Z. Li**, **W. Liu**, and **L. Yuan** (2023). Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment.

190. **Zhu, T.**, **X. Wang**, **X. Qin**, and **M. Li**, Source tracing: Detecting voice spoofing. *In Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2022.