Speech Emotion Recognition with Application to Mental Health: A Tensor Perspective

Sandeep Kumar Pandey

Speech Emotion Recognition with Application to Mental Health: A Tensor Perspective

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

Sandeep Kumar Pandey

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

February 2022

Certificate

This is to certify that the thesis entitled "**SPEECH EMOTION RECOGNITION WITH APPLICATION TO MENTAL HEALTH: A TENSOR PERSPECTIVE**", submitted by **SANDEEP KUMAR PANDEY** (156302006), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:
Guwahati.

Dr. Hanumant Singh Shekhawat
Asst. Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

Dated:
Dharwad.

Prof. S. R. Mahadeva Prasanna
Professor
Dept. of Electrical Engg.
Indian Institute of Technology Dharwad
Dharwad - 580 011, Karnataka, India.

To

My parents

who are everything in my life

My nephews **Arjun** and **Kabir**

for their unconditional love

My **Tuffy**

who is blessing me from heavens

Acknowledgements

I am obliged to Goddess Kamakhya for showing the light whenever there was darkness. This thesis is entirely dedicated to the lotus feet of the mother goddess.

This thesis would not have been possible without the immense help and support of several people in various measures. I take this opportunity to convey my most sincere acknowledgment to all of them. I would like to express my deepest and most heartfelt gratitude to my supervisors Prof. S.R. Mahadeva Prasanna and Dr. Hanumant Singh Shekhawat, for their constant guidance, help, and encouragement throughout my research work. Their insightful feedbacks have helped me much in improving the quality of my thesis. More than guiding my thesis work, they have shaped my personal and professional life through their mentorship. Without them, I would not have been where I am today. Notably, the weekly meetings with Prof.S. R. Mahadeva Prasanna acted as the periodic excitation to my research work and helped me to upgrade as a researcher continuously. Moreover, the ease of discussion with Dr. H.S.Shekhawat encouraged me to clear any doubts whenever it aroused. I am grateful to Prof. Samarendra Dandapat, the Chairman of the Doctoral Committee for providing valuable suggestions on my work throughout the years. I want to express my sincere gratitude to Prof. Rohit Sinha and Dr. Tony Jacob, distinguished members of the Doctoral Committee for their constructive criticisms during my progress seminars.

I would also like to convey my gratitude to all technical and non-technical staffs of the EEE department for their help and support throughout my Ph.D. duration; especially, I acknowledge Mukut Da for timely forwarding of different applications.

I sincerely thank to Prof. Ravi Jasuja of Harvard Medical School for his valuable suggestions on my research work related to mental health diagnosis. His invaluable suggestions helped me to view the experiments from a medical perspective as well as his constant guidance in writing research paper with medical phonology shaped up my writing skills to a huge extent.

I am thankful to my friend Shikha Baghel for her assistance in writing my thesis. I would also take this opportunity to thank Saswati Rabha and Sukanya Biswas for their constant support and encouragement throughout my PHD journey.

Lab environment matters a lot when it comes to research. I would like to thank my labmates - Shoubhik, Mrinmoy, Anik, Moakala, Shikha, Saswati and Brij for keeping the environment jovial and

cheerful.

I attribute this achievement to my parents,brother and sister and brother-in-law for their constant blessings, support, silent prayers for my success and moreover, making me stand in this position.

Sandeep Kumar Pandey

Abstract

Speech Emotion Recognition (SER) has been an active area of research ever since the need for smooth and natural Human-Computer Interaction (HCI) came into play. This thesis aims to develop an SER system based on an amalgamation of Tensor Factorization and Neural Network-based learning to mitigate several issues while using contemporary deep learning architectures. This, in turn, is helpful towards recognizing the mental health issues such as depression, anxiety, etc., from speech signals as it is shown in the literature that mental health and emotions are highly correlated. As such, this thesis tries to provide techniques to incorporate emotional information to assess mental health conditions from speech signals, thereby helping the psychologists assign a depression score to patients based on their experience and machine-generated score, thereby mitigating any human bias which might creep in human-only situations.

In the first work, several tensor-based architectures are explored for the task of Speech Emotion Recognition. A tensor Attention Layer is proposed, which helps to focus on class-specific regions of the speech mel-spectrograms and provides emotion-focused inputs to the Tensor Factorized Neural Network. A 3D AG-TFNN is proposed to leverage multi-dimensional information from 3D mel-spectrogram tensors, incorporating delta and double-delta information along the third mode of the input tensor. A parallel AG-TFNN is proposed to leverage complementary information from mel-spectrograms and modulation spectrograms and fused using a 3D tensor. Experimental evaluation on the state-of-the-art datasets such as Emo-DB and IEMOCAP demonstrates the effectiveness of the proposed approaches over the baseline CNN+LSTM architecture, with the added advantage of less computational complexity and a simpler architecture.

The second work delves into the domain of multi-cultural SER by focusing on the two aspects - universality and cultural specificity of emotions. Thus, we propose two methods, one incorporating cultural specificity and another demonstrating the universal nature of emotions across cultures. In the first method, we propose a novel technique to make a multi-cultural SER by incorporating impactful factors such as speaker and language as markers of cultural distinctiveness. We develop a language and a speaker model to get language and speaker embeddings, and a multi-modal fusion architecture is proposed to fuse the information along with emotional cues. Moreover, in the second method, a triplet-loss-based multi-cultural SER is proposed, which tries to normalize speaker and cultural variabilities and focuses on learning emotions, irrespective of culture. Experiments conducted on a collection of

five language emotion datasets show the proposed technique's robustness in predicting emotions in a leave-one-language-out setting. The system's design allows for incorporating a new language and speaker without needing to retrain the whole system again.

In the third work, we proposed a tensor-based architecture for the task of Multiple Instance Learning when a bag of utterances for a speaker is available, and inferences about the speaker label must be drawn using the bag of utterances. The conventional MIL architectures, such as the baseline CNN-MIL system, suffer from the inherent drawbacks of not considering relative and shared information across the utterances in a bag. These techniques rely on inferring labels for individual utterances and averaging or max-pooling the labels to infer the speaker-level labels. The tensor-based architectures solve this problem by considering the utterances as the third mode in addition to the time and frequency modes in speech spectrograms. As such, TFNNs, by their rich mathematical framework, try to capture the shared information across the utterances of a bag by tensor factorization where the input tensor is projected over three subspaces - time subspace, frequency subspace, and utterance subspace. This helps to utilize the shared information and generate a single speaker/bag level probability for the specified task. To this end, we proposed two tensor MIL architectures - 3D TFNN and 3D TFNN+Attention. Comparison with the state-of-the-art proves that both the proposed techniques effectively capture depression-related information across bags of utterances. Moreover, additional analysis on the optimal number of utterances per bag is also presented to shed light on the model performance when using varying bag sizes.

In the last work, we propose emotion information fusion using Tensor-based fusion approaches for depression classification. Since emotions in speech are highly affected due to an individual's underlying mental health issues, it becomes highly relevant when it comes to automatic assessment of clinical depression using speech. Two fusion approach is explored - Inner-Product based fusion and Elementwise Weighting. The emotion embedding tensors to be fused are generated using pre-trained TFNNs on six English SER datasets since the Depression dataset is also English. Moreover, a multi-modal approach is also explored using audio and text modalities. BERT-based embeddings are explored for text transcripts and fused with audio embeddings learned from mel-spectrogram representations. Two fusion approaches are explored - Late feature fusion and Score Fusion.

The major contributions of the current thesis are as follows:

- A tensor attention layer to provide emotion focused tensor inputs to TFNN.

- Parallel AG-TFNN to leverage complementary information from two speech representations.
- A Multi-lingual Emotion classification system that incorporates language and speaker embeddings along with emotion embeddings to adapt to new culture and speakers.
- A triplet-loss based multi-lingual architecture which tries to normalize language and speaker variabilities, thereby providing a more general solution to cross-cultural adaptivity of SER systems.
- A Tensor factorization based Multiple Instance Learning (MIL) architecture along with utterance level attention and statistics pooling for Depression classification from speech signals.
- Emotion information fusion to aid depression diagnosis using two tensor-fusion approaches - weighted fusion and inner-product-based fusion. Multi-task and Multi-modal architectures are also explored to exploit text-based sentiment embeddings to aid depression diagnosis.

Keywords: Speech Emotion Recognition, Deep Learning, Tensor Factorization, Mental Health, Depression Diagnosis, Multi-cultural, Fusion, Multi-modal, Multi-task

Contents

List of Figures

List of Tables

List of Acronyms

AF	Articulatory Features
AG-TFNN	Attention-Gated Tensor Factorized Neural Network
ALS	Alternating Least Square
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representation from Transformers
BSS	Blind Source Separation
CNN	Convolutional Neural Network
CP	CANDECOMP/PARAFAC
DNN	Deep Neural Network
DP	Double Projection
DTNN	Deep Tensor Neural Network
EEG	Electroencephalogram
ELM	Extreme Learning Machine
ELU	Exponential Linear Unit
EM	Expectation-Maximization
EMD	Empirical Mode Decomposition
EP	Emotion Profile
EVC	Eigen Voice Conversion
FC	Fully Connected
FF	Feed Forward
GABA	Gamma-Amino Butyric Acid
GeMAPs	The Geneva Minimalistic Acoustic Parameter Set
GMM	Gaussian Mixture Model

GRU	Gated Recurrent Unit
HCI	Human Computer Interaction
HMM	Hidden Markov Model
HOSVD	Higher Order Singular Value Decomposition
IEMOCAP	The Interactive Emotional Dyadic Motion Capture Dataset
KELM	Kernel Extreme Learning Machine
LDA	Linear Discriminant Analysis
LFLB	Local Feature Learning Blocks
LLD	Low Level Descriptors
LOSO	Leave One Speaker Out
LSTM	Long-Short Term Memory
MAP	Maximum Aposteriori
MFCC	Mel Frequency Cepstral Coefficients
MIL	Multiple Instance Learning
MPCA	Multilinear Principal Component Analysis Network
NMF	Non-negative Matrix Factorization
PARAFAC	Parallel Factors
PCA	Principal Component Analysis
PHQ	Patient Health Questionnaire
RBM	Restricted Boltzmann Machine
RMS	Root Mean Square
RNN	Recurrent Neural Network
SD	Speaker-dependent
SER	Speech Emotion Recognition
SI	Speaker-Independent
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
TFNN	Tensor Factorized Neural Network
T-sne	t-distributed Stochastic Neighbor Embedding
TT	Tensor Train

UA/UAR	Unweighted Accuracy
UBM	Universal Background Model
UBSS	Underdetermined Blind Source Separation
VMD	Variational Mode Decomposition
WA/WAR	Weighted Accuracy
W.H.O	World Health Organization

1

Introduction

Objective

Speech emotion is one of the essential paralinguistic aspects of the speech signal. It provides naturalness to speech and thus is one of the most sought-after factors when it comes to improving Human-Computer Interaction (HCI) based systems such as automation of call-centers, voice-assistants such as Alexa, Siri, Cortana, etc. The goal to correctly identify an utterance's emotional state becomes significant to receive a proper emotion-motivated reply from HCI systems. However, speech emotion recognition is a challenging problem because of several variabilities. Emotions are often debated to be culturally specific or culturally universal. As such, the cultural variability factors comes into play which restricts the performance in the multi-cultural scenario. Moreover, other factors, such as speaker independence, text independence, etc., also play a significant role in Speech Emotion Recognition(SER). This thesis aims to identify such variabilities in multi-lingual scenarios and propose robust deep learning-based network architectures to recognize the emotional states of utterances, even when the language is unseen to the trained system. A tensor-based approach is proposed to model the emotions from tensors formed out of speech representations, as tensors provide an intuitive view into the operation of 2D/3D speech representations. Also, tensor-based architectures provide several advantages over contemporary CNN-based architectures. Moreover, a direct implication of speech emotion is to aid automatic assessment of the mental health of individuals. To this end, the thesis proposes to automatically detect depression from speech utterances and correlate it with the emotional states of the utterances.

1.1 Introduction

With the advent of technology, there is an ever-rising need to make Human-Computer Interaction (HCI) more natural. As speech is the easiest and most effective form of communication for human beings, it is natural that we would want our machines to be able to understand us completely based on our speech commands. However, apart from the message, speech signal carries secondary-level information also, such as gender, emotional states, etc. Therefore, it becomes evident that the HCI should also be able to capture that information and use it to make the interaction process smoother and contextual pertaining to the speaker's emotional state. Therefore, speech emotion recognition (SER), which aims to identify the emotional state from speech utterances, has drawn particular interest among researchers. Moreover, applications of SER have also been seen in mental health analysis [1] [2], call

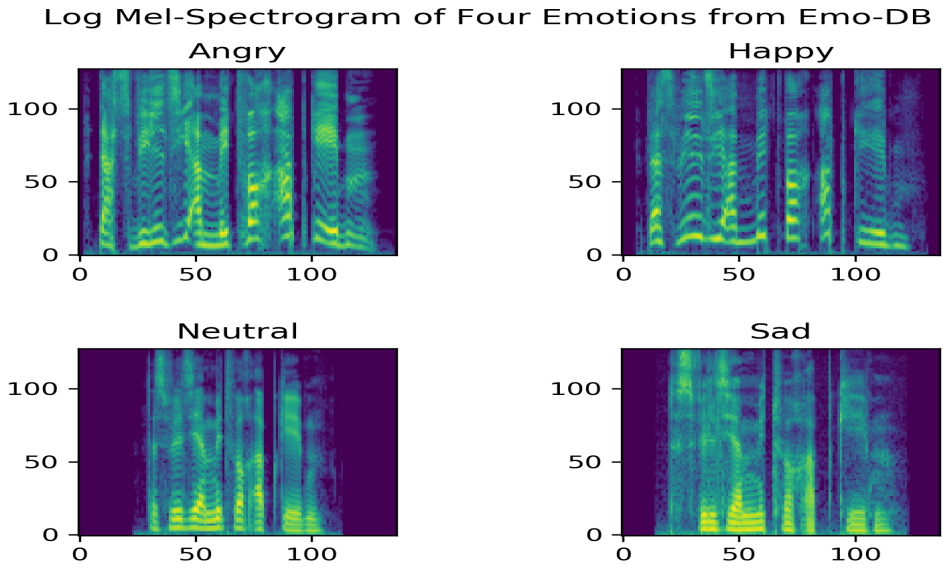


Figure 1.1: Log Mel Spectrogram generated from utterances of same male speakers exhibiting different emotions while uttering the same sentence from Emo-DB Dataset.

centers, intelligent robots, etc., which has further caught the fascination of researchers to explore SER. As such, emotion recognition has surfaced as an ever-emerging research area.

Emotions in speech is manifested as long-term contextual dependencies in the form of variations in energy of frequency content over time. As such, various emotions exhibit energy concentration in different frequency regions in the acoustic frequency range. Emotions such as Angry and Happy exhibit higher energy concentration toward the high-frequency regions. However, Sadness is marked by more presence of energy toward lower frequency range. Neutral emotion is often characterized as evenly distributed energy content over the entire acoustic frequency range [3].

The way utterances are labeled for underlying emotional states also plays a significant role in SER studies. Two types of labeling are explored in literature based on different viewpoints - Dimensional Labels and Categorical Labels. Based on the *Circumplex Model* [4], emotions are broadly categorized in two-dimensional categories -

- *Arousal* - based on the activation\deactivation. It refers to the amount of energy required to express a certain emotion. Emotions such as Joy, Anger, and Fear often arouse the sympathetic nervous system, which results in an increase in blood pressure, heart rate, and changes in respiratory movements, resulting in a speech characterized by loudness, fastness, and higher average pitch and a wider pitch-range. The opposite happens for emotions such as Sadness, with

an increase in salivation, resulting in a speech characterized by slowness, lower high-frequency energy, and low average pitch value.

- *Valence* - based on pleasant\unpleasant characteristics. It refers to the difference in affect between emotions. For example, Joy and Anger both belong to the high-arousal category but convey different affect. Joy belongs to pleasant affect, while Anger belongs to unpleasant affect. The valence dimension characterizes this phenomenon. However, there is no consensus on which acoustic parameters correlate with valence.

However, categorical labels for emotions are motivated by *Palette Theory* [5], which states that any emotion can be decomposed into primary emotions - Joy, Sadness, Anger, Neutral, Fear, and Surprise, similar to colors which can be decomposed into primary colors. Nevertheless, there is always an ambiguity regarding the categorical labeling of emotions. An utterance that might sound Neutral to one individual may be perceived as Happy by others. This ambiguity makes the task of SER for categorical labels a challenging one.

1.2 Emotion recognition from speech

Speech signal bears information on two levels. On the primary level, it conveys information such as message and speaker identity, and at the secondary level, information such as emotional content, naturalness, intelligibility, etc. Emotions bring naturalness to speech. In an era of human-machine interaction (HCI), with the development of intelligent dialogue systems and voice assistants such as Alexa, Cortana, Siri, and Google Assistant, the need to understand the intentions behind a dialog becomes significant to deliver emotion-aligned automatic replies by the system. Emotions being a significant marker behind intentions in an utterance [6], the effort is more on making the machine-generated speech sound more naturalistic. Moreover, applications of SER have also been seen in mental health analysis [1] [2], call centers, intelligent robots, etc., which has further caught the fascination of researchers to explore SER. As such, emotion recognition has surfaced as an ever-emerging research area.

Emotions in speech are affected by several factors such as speaker identity, recording environment, cultural variations, gender, mental state, etc. As such, to realize a robust SER system capable of handling multiple variabilities, the design should incorporate features or architectures that either adapt to new variabilities or can generalize across the variabilities [7], [8]. Moreover, mental health

issues such as depression, anxiety, stress, dementia, etc., directly affect the emotional states underlying a speech utterance. Hence, SER can be considered a precursor to mental health diagnosis using behavioral signals such as speech. Moreover, by conducting a bias-free diagnosis of depression from speech, SER based depression diagnosis system provides a promising aid to psychologists to augment their experience-based scores with machine-generated ones.

1.3 Traditional approaches for SER

Feature representation is one of the vital aspects to be considered in any emotion recognition study. Distinct emotion categories have different inherent properties and as such, portraying them using a single feature type is quite a difficult task. As such, some different combination of feature types has been explored in literature, such as prosody, voice quality, and spectral features [6]. Traditional SER methods explored auditory features such as mel-frequency cepstral coefficients(MFCCs), F0, energy, formant based features in combination with a classifier such as Gaussian Mixture Model (GMM) [9], Hidden Markov Model (HMM) [10], Support Vector Machine (SVM) [11] etc to predict labels for emotional utterances. All these traditional techniques were based on the auditory perception of emotions in speech and prior knowledge of hand-crafted features. Also, feature sets defined for SER challenges such as IS09 feature set of The INTERSPEECH 2009 Emotion Challenge [12], IS13-ComParE/IS17-ComParE feature sets [13], [14], GeMAPs [15] feature set served as the standard baseline feature sets till now for SER research. These feature sets comprised low-level descriptors (LLDs) extracted from frames of speech signals and utterance level representations obtained by applying statistical functionals to the frame-based LLDs. The inherent issue with all such features was the use of frame-based feature extraction, which ignores the point that the emotional state underlying an utterance is spread over the entire length. As such, features that capture this temporal dependence of emotions need to be investigated further. Jeong et al. [16] investigated the combination of spectral and prosody features on a GMM framework for four emotion classes of EMO-DB. The accuracy reported was 85% with more anger versus happiness confusion. Lugger and Yang [17] used prosody and voice quality features on seven emotion classes on the GMM framework, but the classification did not improve. Sun and Moore [18] used a combination of spectral, prosody, and voice quality features, and SVM's were trained with it. The reported accuracy for EMO-DB was 80%.

The shortcomings of the short-term spectral features led to the exploration of long-term spec-

tral features. Wu et al. [19] have proposed a new feature representation capturing the long-term spectro-temporal modulations of a speech utterance, called modulation spectral features. This led to the exploration of modulation spectral features in speaker identification, speech recognition, speaker diarization etc [20] [21] [22].

1.4 Deep Learning approaches for SER

As deep learning field gained momentum, time-frequency representations of speech gained popularity for feature extraction and classification in an end-to-end framework [23], [24], [25]. Mel-frequency spectrograms emerged as the best choice for feature when deep learning classifiers are used for SER task [26], [27], [28]. Moreover, works in [29], [30], [31] and [26] have utilised the raw speech directly for feature extraction and classification, thereby eliminating the need of handcrafted features.

With the advent of the deep learning era, fields such as image and speech processing saw a surge in applications with high recognition performances. The baton of feature extraction is passed on from hand-crafted traditional features to automatic deep features extracted using deep neural network architectures. The popularity of Convolutional Neural Networks (CNN) in visual recognition tasks prompted speech researchers to leverage it for feature extraction in speech processing. In works [24], [32], [33] a deep CNN network was utilized to extract features from speech spectrograms and classify it into emotion labels. However, CNNs are known to extract local features, not considering in view the temporal aspect of emotions. As such, to explore the temporal context in utterances, diverse Recurrent Neural Network (RNN) architectures such as LSTMs, GRUs, etc., are widely explored in SER domain [34], [35]. Moreover, to exploit the usefulness of both CNNs and RNNs, combined CNN+LSTM architectures are now a trend in the SER field, with CNNs as feature extractors and LSTMs for temporal dependency modeling followed by a fully-connected (FC) layer for classification [26], [28], [36].

1.5 Challenges in SER

Speech being the most preferred form of communication, researchers have focused on making Human-Computer Interaction (HCI) such as automated call-center, voice assistants such as Alexa, Cortana, Siri, etc., sound more natural and provide emotion-aligned responses to the user. However, the task of emotion recognition from the speech is not an easy one. Due to a large number of

variabilities involved, both inter-class and intra-class, SER poses a tough challenge to the researchers. Some of the research challenges corresponding to SER in general are -

(i) It is not clear which speech feature corresponds or is prominent in which emotion category due to the highly variable nature of speech emotion.

(ii) Some of the variabilities that impact the emotional state underlying an utterance are speaker-specificity, speaking rate, style, linguistic content, etc. These variabilities directly impact the prominent speech features based on pitch and energy contours [37]. This makes two utterances belonging to the same emotion classes quite apart in feature space.

(iii) Moreover, there may be more than one perceived emotion in the same utterance; each emotion corresponds to a different portion of the spoken utterance. Also, defining boundaries between emotions in an utterance is a difficult task.

(iv) Another vital factor that imparts variabilities in emotions is the cultural specificity of emotions. Many works have shown that emotions are cultural-dependent up to a large extent. This makes the generalization of SER systems to an unseen language difficult.

(v) Also, the emotional datasets available for research purposes are mostly acted emotions. As such, the utterances' emotions may not match real-life emotions. This poses a tough challenge for SER models to achieve high performance on real-world data as acted data often have exaggerated emotions.

Numerous other challenges are associated with SER, such as data availability in low-resource languages, real-life emotion versus acted emotion data, etc. However, the core challenge lies in generalizing SER models to perform stably across the speaker, linguistic and cultural variabilities.

1.6 Motivation for the present work

To account for the challenges posed by SER, several studies in the literature have adopted techniques such as cross-corpus SER, Leave-one-speaker-out (LOSO) validation technique, etc., using both conventional and deep learning approaches. However, using complex deep learning architectures becomes a limitation when the size of the dataset at hand is small, which is often the scenario for SER datasets. As such, the motivation for the work presented in this thesis is listed below -

(i) Deep Learning Techniques such as DNN and CNN+LSTM-based models have shown considerable success in the SER domain. However, a major limitation associated with such deep learning techniques is the increase in parameter size with more complex and deeper architectures. A considerable amount of data is required to train such deep architectures effectively without over-fitting/under-fitting issues, which is not feasible in all scenarios. As such, there is a need for a lightweight and equivalently effective architecture which can extract discriminative features from speech representations, without increasing the burden of parameters with deeper architectures.

(ii) Moreover, one of the significant challenges in the SER domain is the cross-cultural generalization of SER systems, which limits the application of SER in several real-world situations. Also, within a specific culture, the major variability associated with emotions is speaker specificity. To solve this challenge, SER systems are required, which can either adapt themselves to new speakers and languages on the go or can generalize emotional information across language and speaker variabilities.

(iii) One of the critical aspects of emotions in a speech utterance is how it gets affected due to the mental state of an individual. Several mental health issues influence the underlying emotional state in a speech utterance. As such, emotional information in an utterance can be exploited to effectively diagnose such mental health issues and can be used as an additional aid to clinicians.

This thesis aims to approach solving the challenges mentioned above and provide some novel insight into the SER domain.

1.7 Organization of the thesis

- Chapter 2 provides a basic introduction to tensor algebra required for comprehending the various Tensor-based techniques used in this thesis. It provides an overview of the standard terminologies and notations along with two standard tensor factorization techniques popular in literature.
- Chapter 3 reviews the Tensor-based methods employed in the speech processing domain. The chapter discusses methods based on tensor factorizations for speech processing research, such as speaker recognition, speech recognition, blind source separation, etc.
- Chapter 4 introduces the Attention-Gated Tensor Factorized Neural Network, which is shown to be an effective alternative to CNN+LSTM-based SER systems. Standard speech representations

such as 2D and 3D Mel-Spectrogram and Temporal Modulation Spectrogram are explored to investigate the emotion salient information capturing effectiveness of the Tensor Factorization-based architectures. The hidden layers are explained as Deep Tensor Factorization based on the Tucker Decomposition but with a unified discriminative objective function to learn the factor matrices in a discriminative sense. The core tensor produced in each hidden layer is the feature associated with that factorization layer. Mel Spectrograms are naturally in 2D tensor form. Thus TFNN and AG-TFNN become an appropriate choice over baselines such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) by providing reduced parameters to be learned and simple architecture. Experiments conducted on standard emotional speech datasets- Emo-DB and IEMOCAP show that TFNN and AG-TFNN surpass the state-of-the-art CNN+LSTM combination with fewer parameters.

- Chapter 5 explores more general and basic problems related to SER - cultural and speaker specificity of emotions. Both universality and cultural specificity of emotions are debated in the literature. Thus we propose two methods, one incorporating cultural specificity and another demonstrating the universal nature of emotions across cultures. In the first method, we propose a novel technique to make a multi-cultural SER system by incorporating impactful factors such as speaker and language as markers of cultural distinctiveness. We develop a language and a speaker model to get language and speaker embeddings, and a multi-modal fusion architecture is proposed to fuse the information along with emotion cues. Moreover, a triplet-loss-based multi-cultural SER is proposed, which tries to normalize speaker and cultural variabilities and focuses on learning emotions, irrespective of culture. Experiments conducted on a collection of five language emotion dataset shows the robustness of the proposed technique in predicting emotions in leave-one -language-out setting. The system's design allows for the incorporation of a new language and speaker without retraining the whole system again.

- Chapter 6 explores the effectiveness of Tensor-based methods in analyzing and diagnosing mental health issues such as Clinical Depression. Tensor Factorization-based Multiple Instance Learning (MIL) is proposed, which shows significant improvement over the conventional CNN-based MIL systems. It even surpasses state-of-the-art techniques for DAIC-WOZ Depression dataset. Moreover, an utterance level attention is proposed along with statistics pooling to decrease class confusion in the Tensor MIL framework.

- In chapter 7, the effect of emotional information on clinical depression diagnosis is explored. Two Tensor-based fusion approaches - weighted fusion and inner-product-based fusion are proposed to effectively incorporate emotional information for depression classification in a parallel-network framework. Text modality is also explored to aid depression diagnosis using multi-task framework. A preliminary investigation into multi-modal architecture using CNN and BERT is also explored.

- In chapter 8, a detailed summary of the present work is reported by highlighting the significant contributions made in this thesis. A discussion regarding the issues related to the present work is also included in this chapter. Finally, the directions for future work related to this thesis are discussed.

2

Tensors Preliminaries

Objective

Tensors are powerful mathematical objects capable of handling multi-modal data naturally, which comes in multiple dimensions. Tensors are the higher-order generalization of matrices. It has the inherent property of capturing natural structure, latent semantic information, and high-order interactions. With the surge in deep learning, many applications require a massive amount of structured high-dimensional data. The vectorization of such data poses a challenge regarding data requirement and computational complexity. This chapter introduces the fundamentals of tensors needed to understand tensor-based techniques. The various notations used to depict tensor objects, the tensor operations, tensorization and matricization, tensor projections, and decompositions are discussed in detail.

2.1 Introduction

Tensor was first introduced by William Ron Hamilton in 1846 and later got introduced to the scientific community through the book The Absolute Differential Calculus [38]. Tensors are generalizations of multi-dimensional arrays. Due to its inherent structured representation of data and the capability to alleviate complexity of multi-dimensional arrays, tensors have been applied in a wide variety of research ranging from tensor data completion [39], tensor classification/regression [40], blind-source separation [41], noise reduction [42] etc. A tensor representation captures complex interactions among input features that would not be evident on flattened data [43]. However, tensors are also affected by the curse of dimensionality when the size of tensors increases. To counter this issue, tensor decomposition comes into play which keeps the tensor form intact while reducing the dimensionality and keeping the relevant information only.

2.2 Tensor Unfolding and Multilinear Projection

A tensor is a multi-dimensional or multi-index numerical array [44] e.g $\mathcal{X} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_n \times \cdots \times I_N}$, where $n \in 1, 2, \cdots, N$ denotes the modes of the tensor which corresponds to time, frequency, trials, utterances etc [45]. The number of modes in a tensor corresponds to the order of the tensor. As such, tensors provide a realistic representation of many multi-variate data objects evolving over multiple independent variables [44] [45]. Tensor manipulation generally requires reordering the elements of a tensor in matrix form, which is done using mode-n unfolding. For a nth order tensor $\mathcal{X} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_n \times \cdots \times I_N}$, mode-$n$ unfolding is achieved by arranging the mode-n fibers of $\mathcal{X}$ as columns

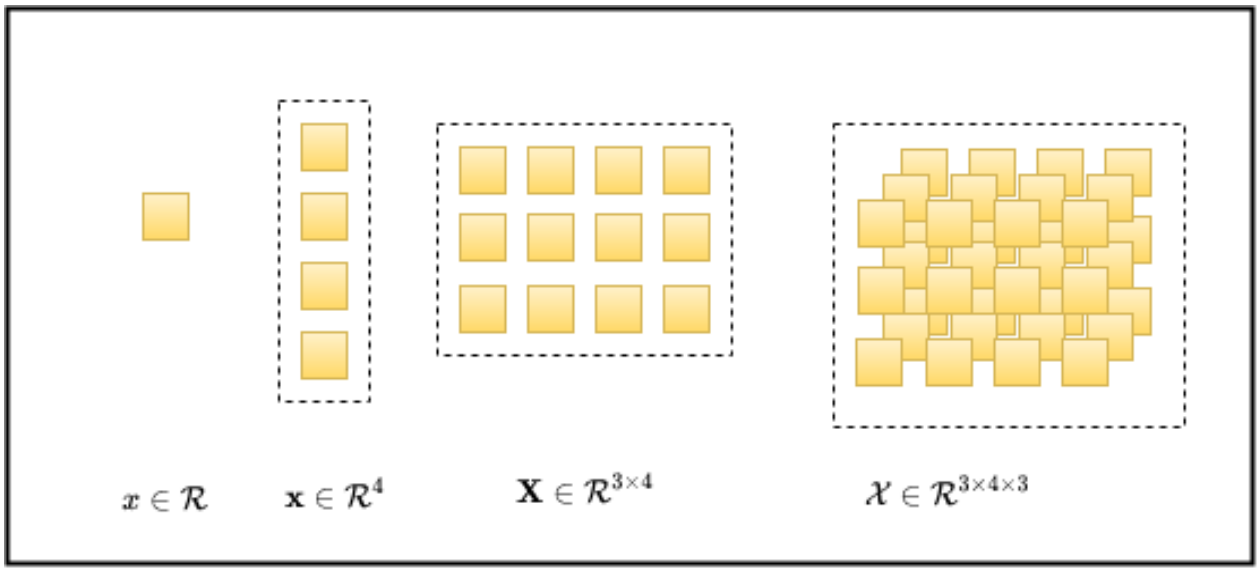


Figure 2.1: Different order Tensors. A scalar 'x' is represented by a zero-order Tensor, a vector '$\mathbf{x}$ ' is represented by a first-order tensor, a matrix '$\mathbf{X}$' is represented by a second-order tensor, and a third-order tensor '$\mathcal{X}$' is of the form of a cuboid.

of the resultant matrix [44]. It is denoted by $\mathbf{X}_{(n)} \in \mathbf{R}^{I_n \times l}$ where $l = (I_1 \times \cdots \times I_{n-1} \times I_{n+1} \times \cdots \times I_N)$. As such, an Nth order tensor yields N number of mode-n matrices, one corresponding to each mode.

The mode-n product of a tensor $\mathcal{X} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_n \times \cdots \times I_N}$ with a matrix $\mathbf{U} \in \mathcal{R}^{J_n \times I_n}$ results in a tensor $\mathcal{Y} \in \mathcal{R}^{I_1 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N}$, and is mathematically denoted as,

$$\mathcal{Y} = \mathcal{X} \times_n \mathbf{U} \tag{2.1}$$

where, $\times_n$ denotes the mode-n product and $n = 0, 1, \cdots N$. Each of the elements of $\mathcal{Y}$ can be seen as a sum of products of the corresponding elements in $\mathcal{X}$ and $\mathbf{U}$. Using the mode-n unfolding, the mode-n product can be written as-

$$\mathbf{Y}_{(n)} = \mathbf{U}\mathbf{X}_{(n)} \tag{2.2}$$

where, $\mathbf{Y}_{(n)}$ and $\mathbf{X}_{(n)}$ are the mode-n unfolded matrices of the tensors $\mathcal{Y}$ and $\mathcal{X}$ respectively.

Multilinear subspace learning requires understanding multilinear projections, which is motivated by the fact that a tensor of a higher dimension may reside in a tensor subspace of lower dimensions. As such, considering the general case of a Nth order tensor $\mathcal{X} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_n \times \cdots \times I_N}$, the tensor space consists of the outer product of N vector spaces $\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_N$ and is denoted by $\mathbf{R}^{I_1} \otimes \mathbf{R}^{I_2} \otimes \cdots \otimes \mathbf{R}^{I_N}$. Thus, a tensor $\mathcal{X} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_n \times \cdots \times I_N}$ can be projected to a lower dimensional tensor

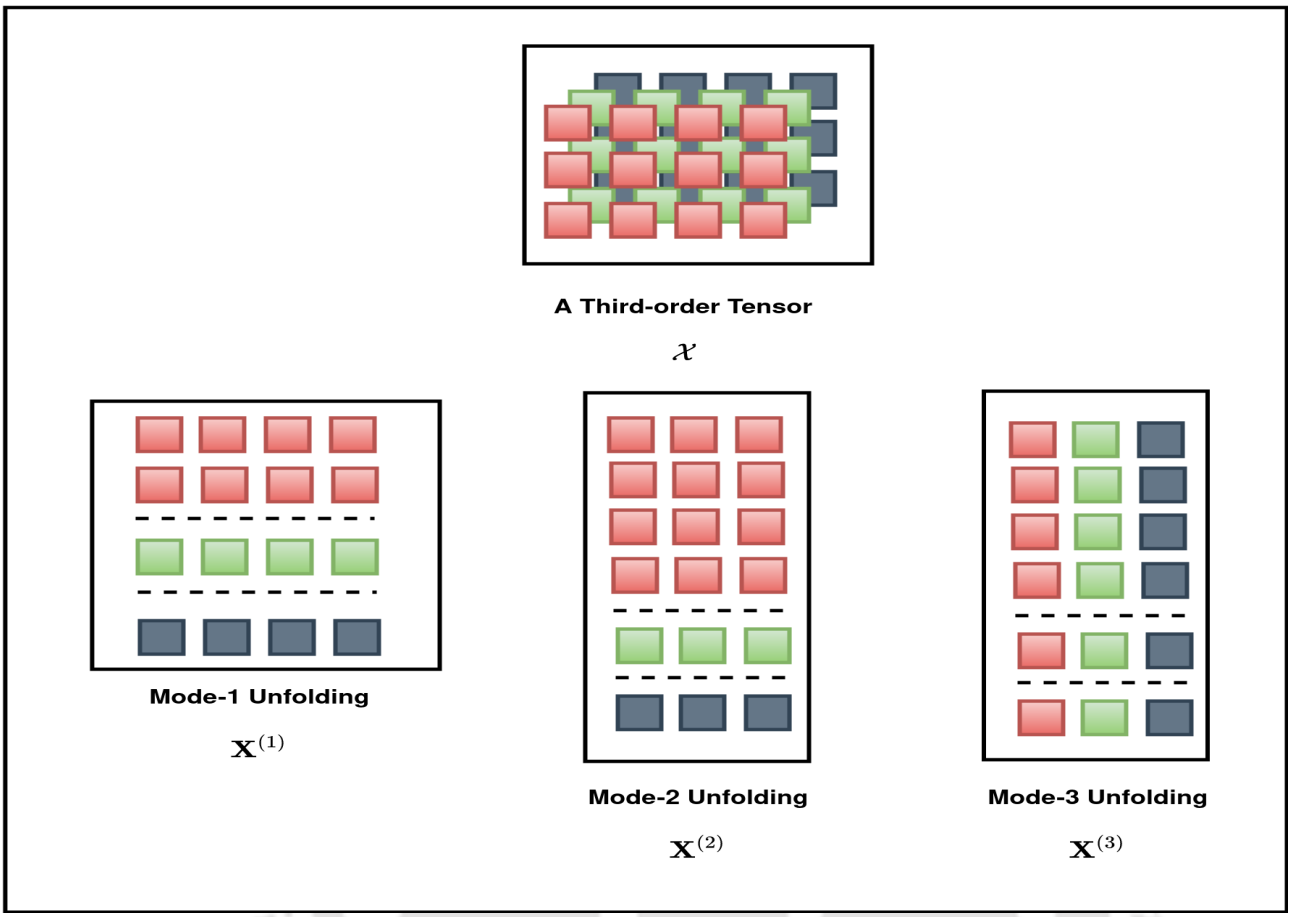


Figure 2.2: Unfolding of third-order tensor $\mathcal{X}$ along the three modes generating three unfolded second-order tensors (matrices) - $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}$.

$\mathcal{Y} \in \mathcal{R}^{J_1 \times J_2 \times \cdots \times J_n \times \cdots \times J_N}$ residing in the subspace $\mathbf{R}^{J_1} \otimes \mathbf{R}^{J_2} \otimes \cdots \otimes \mathbf{R}^{J_N}$ where $J_n \leq I_n$, using N projection matrices $\left\{\mathbf{U}^{(n)} \in \mathbf{R}^{J_n \times I_n}, n = 1, 2, \cdots, N\right\}$, one corresponding to each mode of the tensor $\mathcal{X}$ and is denoted by,

$$\mathcal{Y} = \mathcal{X} \times_1 \mathbf{U}^{(1)^\dagger} \times_2 \mathbf{U}^{(2)^\dagger} \times_3 \cdots \times_N \mathbf{U}^{(N)^\dagger} \tag{2.3}$$

where, $\mathbf{U}^{(n)^\dagger}$ denotes the pseudoinverse of $\mathbf{U}^{(n)}$.

2.2.1 Inner Product

The inner product of a tensor $\mathcal{X} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_n \times \cdots \times I_N}$ with another tensor of same dimensions $\mathcal{Y} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_n \times \cdots \times I_N}$, is the sum of the product of their entries, i.e -

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} x_{i_1, i_2, \cdots, i_N} y_{i_1, i_2, \cdots, i_N} \tag{2.4}$$

2.3 Tensor Decompositions

Tensor decomposition aims to reduce computational complexity while maintaining the inherent data structure. Among the many advantages, the prominent one is tackling the *Curse of Dimensionality.*

2.3.1 CP Decomposition

CANDECOMP/PARAFAC (CP) decomposition has been proposed in numerous fields. However, it boils down to the same concept [46]. The CP decomposition factorizes a tensor into a linear combination of rank one tensors. Mathematically, for an Nth order tensor $\mathcal{X} \in \mathbf{R}^{I_1 \times I_2 \times \cdots \times I_N}$, it can be written as a sum of R rank one tensors -

$$\mathcal{X} = \sum_{r=1}^{R} a_1^r \circ a_2^r \circ \cdots \circ a_N^r \tag{2.5}$$

Here, $\circ$ represents the outer product [44]. It is assumed that the vectors $a_1^r, a_2^r, \cdots, a_N^r$ are normalized. However, for a general case, it can be written as follows.

$$\mathcal{X} \simeq [\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_N] = \sum_{r=1}^{R} \lambda_r a_1^r \circ a_2^r \circ \cdots \circ a_N^r \tag{2.6}$$

where, $\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_N$ are the mode-n factor matrices, such as $\mathbf{A}_1 = [a_1^1, a_1^2, \cdots a_1^R]$, containing R columns with R being the rank of the decomposition. The unit normalization is accounted for by the term λ_r. The decomposition is performed by minimizing an objective function where one of the factor matrices is updated while keeping the others fixed on a rotation basis. This technique is popularly known as *Alternating Least Square (ALS).*

2.3.2 Tucker Decompostion

Tucker Decomposition was first introduced by Tucker in 1963 [47] and was refined in subsequent articles, the most popular and cited one being [48]. It goes by many names such as Three mode factor analysis, Three mode PCA, Higher order SVD (HOSVD), etc. It is a generalization of vector PCA to multilinear arrays. It has found applications in numerous studies such as signal processing [49], Weiner filters [50], TensorFaces for face recognition [51], Human motion [52], Handwritten digits classification [53], etc. Tucker Decomposition factorizes an Nth order tensor into a core tensor having the same order as the original tensor and reduced or equal dimensions with the help of N- Factor

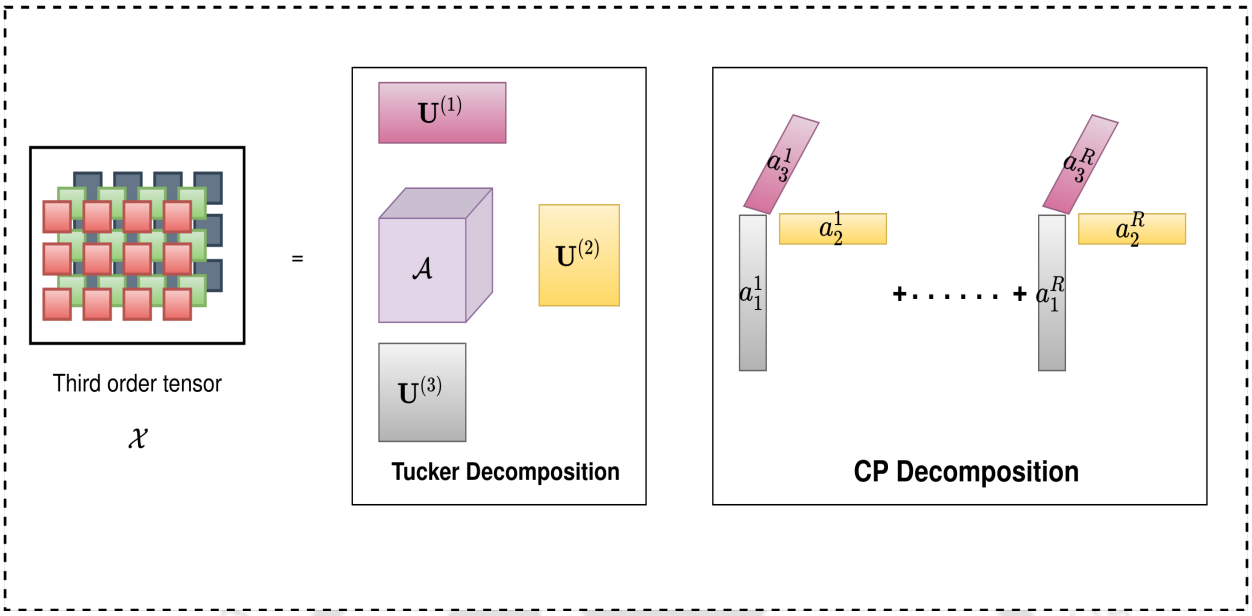


Figure 2.3: Tensor Decomposition of a third-order tensor$\mathcal{X}$

matrices, one for each mode. Mathematically, for an Nth order tensor $\mathcal{X} \in \mathbf{R}^{I_1 \times I_2 \times \cdots \times I_N}$, it can be written as -

$$\mathcal{X} = \mathcal{A} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \cdots \times_N \mathbf{U}^{(N)} \tag{2.7}$$

where, $\mathbf{U}^{(n)}, n \in 1, 2, \cdots, N$ are the factor matrices of dimensions $I_n \times J_n$ having $J_n \leq I_n$ and the core tensor $\mathcal{A}$ can be written as-

$$\mathcal{A} = \mathcal{X} \times_1 \mathbf{U}^{(1)^\dagger} \times_2 \mathbf{U}^{(2)^\dagger} \times_3 \cdots \times_N \mathbf{U}^{(N)^\dagger} \tag{2.8}$$

The major advantages of Tucker Decomposition, which can be utilized in the deep learning domain, are as follows.

- Tucker decomposition breaks down a data tensor into a core tensor with the same number of modes as the original tensor, with a size equal to or less than the original tensor. The core tensor can be used as a feature tensor for subsequent input to classification algorithms. This property is the central theme for the proposed architectures in this thesis. Chapter 4 shows the use of core tensor/ embeddings in classification task. It is also shown that the core tensor features capture emotion information effectively compared to classical features.

- The factor matrices are learned in such a way to extract shared information across all the modes for the given task. Data tensors with multiple modes will generate feature tensors by

multilinear projection with factor matrices, which will have shared information from all the modes. This functionality becomes helpful when dealing with problems such as multiple instances of a speaker's utterance, noise removal from images [54], etc. This property of shared information extraction is utilized in architectures proposed in Chapter 6. The Multiple Instance Learning problem is tackled by exploiting the shared information across multiple utterances of a speaker, stacked along the third dimension of the input tensor.

- Since the size of the core-tensor is a hyper-parameter, this gives us flexibility on the size of the feature we want to retain as per our computational needs. For example, the feature vector generated per utterance using ComParE 2016 feature set is of dimensions 6373 whereas, the dimensions of the flattened embedding depends on the size of the core-tensor. Smaller the core-tensor, the feature vector obtained by vectorizing the core tensor is of smaller dimensions.
- Additional constraints can be imposed on the factor matrices such as non-negativity for image data [55], as image pixel values are non-negative, sparsity on the core-tensor to reduce redundancy, etc.

2.4 Conclusion

Tensors serve as a natural and effective representation when dealing with multi-way/multi-dimensional data, which is often the case in speech processing, image and video procession, deep learning, etc. A brief introduction to various order tensors is given first, which builds the foundation for understanding tensor data and its multi-way components. Processing tensors often requires its matricization to use several matrix algebra techniques. To this end, the mode-n matricization is explained. Furthermore, tensor decomposition for feature extraction/tensor reduction is explained, which is the basis of the proposed approaches in further chapters in the thesis. A comprehensive discussion on the effectiveness of Tucker Decomposition for deep learning approaches is also presented.

3

Tensors in Speech Processing: A Review

Objective

Pattern recognition and deep learning often generate a colossal amount of multi-modal data with high dimensionality, for which Tensors are an intuitive representation. Multilinear algebra, often called Tensor algebra, provides tools such as TUCKER and PARAFAC for feature extraction and classification from high-order datasets, capturing the multi-modal and multi-aspect information present along with the different modalities [40]. The different modalities in data can be time, frequency, trials, session, utterances, channel, noise, etc., which can be captured well using the multi-way factorization tools provided by Tensor algebra. Because of this aspect of tensor factorization to extract multi-way features comprehensively, it has found wide application in image processing, biomedical signal processing, speech processing, etc. The objective of this chapter is to familiarize the reader with various tensor-based techniques used in the speech processing domain. A brief overview of some standard approaches is discussed. A preliminary investigation using speaker recognition as an application is also performed to assert the effectiveness of the conventional tensor-based method.

3.1 Introduction

The dawn of the deep learning era, with improved computational speed and data availability, introduced algorithms to process images, audio, time-series data, etc. The traditional algorithms such as GMM, SVM, HMM, etc., which were used for classification/regression tasks, primarily depended on feature representations in vector form. However, this posed another serious problem - " the curse of dimensionality ". As such, with the increase in vector size, more data is needed to learn decision boundaries effectively. Dimensionality reduction techniques come into play in such scenarios where vector size is large. However, it has its own demerits- the major one being the discarding of information when reducing the dimensions of the feature vectors. To add to this issue, the feature vector extraction process already discarded much information from highly dense and redundant signals such as images, speech, etc. This motivated the researchers to develop techniques that can handle the data in its original form, reduce reliance on hand-crafted features and solve the problem of the high dimensionality of feature vectors.

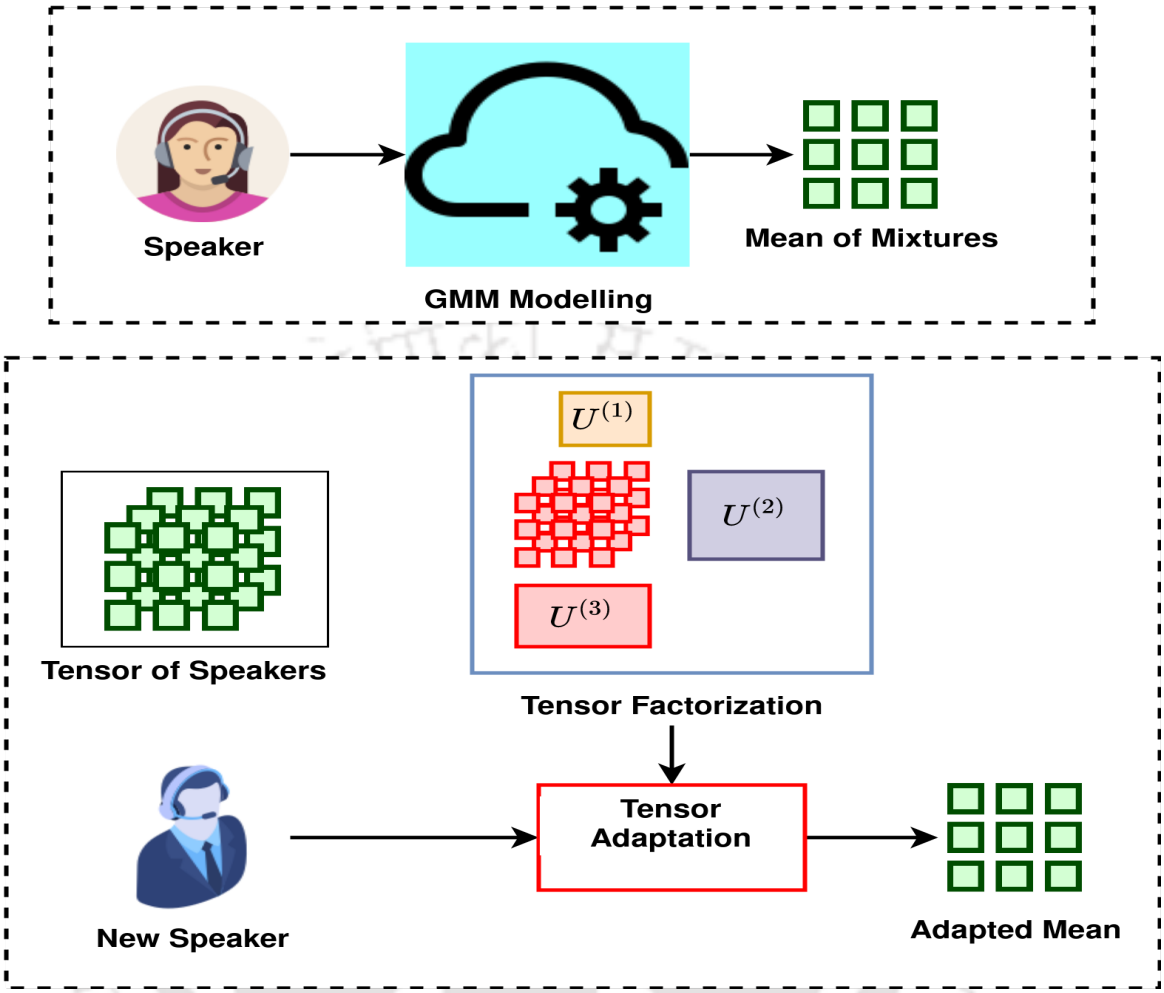


Figure 3.1: Speaker Space Construction using GMM mean matrices of several speakers. Tensor Factorization of the GMM mean tensor of third-order tensor $\mathcal{X}$ along the three modes yields three factor matrices - $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}$.

3.2 Tensors in speech processing

The works described below provides an insight into how tensors can be effectively utilized in speech processing-

3.2.1 Tensors in Automatic Speech Recognition

- In [56], tensors have been explored with emphasis on automatic speech recognition having a large vocabulary. The basic DNN is modified to a novel deep tensor neural network (DTNN) in which one or more layers are double-projection (DP) and tensor layers. The basic idea of the DTNN comes from the motivation and assumption that the underlying factors, such as the spoken words, the speaker identity, noise and channel distortion, and so on, which affect

the observed acoustic signals of speech, can be factorized and be approximately represented as interactions between two nonlinear subspaces. The interactions among these two subspaces and the output neurons are subsequently modeled through a tensor with three-way connections. The conventional DNN and the new DTNN are evaluated on the MNIST handwritten digit recognition task and the SWB phone-call transcription task. The results demonstrate improved performance over conventional DNN.

- To further explore the capability of tensor-based networks in noisy conditions for ASR, [57] explored speech tensors having acoustic and articulatory features as complementary information. The articulatory features (AFs) are extracted using a pre-trained DNN, trained on labeled AF data. Tensor network, similar to the network proposed in [56], is utilized to perform speech recognition in noisy and clean environments. Experimental evaluation demonstrates that placing the Double Projection (DP) layer on top of hidden layers in DNN improves performance in clean and noisy scenarios.

3.2.2 Tensors in Blind Source Separation

Blind Source Separation (BSS) is a technique to estimate the source signals present in a mixture without having any prior knowledge about the source. This method was first introduced in [58] and then continued by researchers to date. Matrix and Tensor Factorization techniques have been extensively deployed for this task and have demonstrated state-of-the-art results. Some of the studies involving the tensor-based method for BSS are discussed in brief below -

- In [59], the task of the convolutive mixture of signals is tackled using PARAFAC decomposition of tensors created through temporal segmentation of multi-channel mixture matrix. Additionally, the parameters of the PARAFAC model are estimated using Alternating Least Square (ALS) optimization and the Majorization algorithm, which tries to minimize the resulting Trace Function. Experiments on synthetically mixed sources in a convolutive fashion demonstrated a correlation greater than 0.70% for the separated source and the original signal.

- In [60], monoaural audio separation is addressed by using tensor factorization with a non-negativity constraint on modulation spectrogram features. The motivation for using modulation spectrograms comes from the view that it highlights redundant patterns in frequencies across similar features, and tensor factorization can efficiently isolate such patterns. The proposed

method was evaluated for seperation of two-source mixtures in unsupervised manner and superior performance was observed compared to standard Non-negative Matrix Factorization (NMF) techniques.

- In [61], Underdetermined Blind Source Separation (UBSS) is tackled using Tensor decomposition and Non-Negative Matix Factorization (NMF) jointly. PARAFAC decomposition is used o estimate the mixing matrix, and NMF is used to estimate the source spectrogram factors. Then Expectation-Maximization (EM) algorithm is used to update the model parameters. Experimental evaluation of synthetic instantaneous and convolutive mixtures of speech and music demonstrates state-of-the-art performance.

3.2.3 Tensors in Speech Enhancement

Speech enhancement techniques are concerned with improving the quality and intelligibility of noisy speech. Earlier work in this area using DNNs has focused chiefly on vector-to-vector regression, extending the capabilities of the single-channel system to a multi-channel system by concatenating the features of multiple channels to form a high dimensional vector. This issue of high dimensionality and the added advantage of exploiting pixel relationships in 2D representations are addressed in recent works using Tensor techniques. A brief overview of the same is presented below.

- In [62], a hybrid architecture is proposed for the speech enhancement task consisting of CNN layers for feature extraction and Tensor Train (TT) Decomposition on the top to reduce model parameters. Experiments performed on multi-channel simulated WSJ0 corpus demonstrate the efficiency of the proposed approach with only 34% of the parameters compared to the CNN-only model.

- In [63], a more flexible approach for controlling the speaker characteristics using the tensor representation of speaker space is described. The proposed method solves the problem of the high dimensionality of supervectors by using the matrix of Gaussian component means in the matrix form. Compared to the EigenVoice Conversion (EVC) method, the proposed approach performed better for less number of adaptation utterances.

- In [64], single-channel speech enhancement is investigated using an LSTM-based network with Tensor Train Decomposition to reduce the size of parameters to be learned. A Time-Frequency

(TF) masking-based approach is utilized, and a Deep Tensornet is proposed with LSTM-TT units as layers. Experimental evaluation on noisy synthetic data created from the clean speech of TIMIT dataset and noise from NOISEX dataset, added on a scale of SNR ranging from $-6DB$ to $+9DB$, in a step of $+3DB$.

3.3 Preliminary Investigation : Speaker Recognition

GMMs serve as a benchmark for the speaker recognition task. Speaker adaptation techniques such as MAP are employed to learn a speaker-dependent model from a pre-trained speaker-independent model. While the GMM-UBM system has proven to be very effective for speaker recognition tasks, the system's performance degrades in a noisy environment and when there is a mismatch between the training and testing conditions [65]. To account for such variabilities, various speaker space algorithms were explored. One such algorithm which is most revered is the Eigenvoice method. It is a model-based speaker adaptation method, working on the intuition of finding a good speaker-dependent model for the new speaker in the space of possible speaker-dependent models. The speaker space algorithms constrain the adapted model of a new speaker to be a linear combination of the basis vectors of the speaker space, which are pre-trained using a wide variety of reference speakers [66]. Eigenvoice algorithm puts an additional constraint that the basis vectors obtained must be orthogonal to capture essential components of variation among the speakers. PCA is a popular method that is used to obtain such basis vectors. The advantage of PCA is that the eigenvectors are arranged according to the magnitude of their contribution to the variance between the reference speakers.

The supervector-based methods opened up a new era of speaker recognition. The Eigenvoice method suffered from the channel and session variabilities too. As such, a new method called Joint Factor Analysis was proposed based on factor analysis, in which, apart from the speaker variabilities, the channel and session variabilities were explicitly modeled by representing the supervector space as a combination of statistically independent speaker and channel spaces [67] [68]. Though the channel factors were specifically meant to model only the channel effects, the experiment conducted in [69] shows that it contains some speaker information too. This gave rise to the idea of using factor analysis as a feature extractor by considering a single space that contains both speaker and channel variabilities, called the Total Variability Space [70]. The speech utterances are represented by a low dimensional vector called i-vector in this total variability space. The i-vector method has become a state-of-the-art

method for speaker recognition experiments.

Most of the "speaker space" methods described above use supervector representation of speakers. Each speaker is represented as a mixture of mean concatenated supervector of very high dimensions. Such a representation fails to distinguish between the mixture and dimension of the mean vector. Also, the dimensionality of the supervectors is very high, and computational load increases as the number of mixtures or dimensions of feature vectors increases. A more natural representation would be where the distinction between mixtures and dimensions is preserved. This is where tensor comes into play. The mixture means matrices corresponding to each speaker-dependent GMM are used to form a third order tensor where the third mode of information is the number of speakers. Tensor, a powerful mathematical modeling tool, can decouple the information along the different modes of variation and can represent them by subspaces along each mode. Using the bases for the subspaces and a suitable adaptation algorithm, speaker adaptation can be performed for new speakers.

3.4 Tensor representation of speaker space

This section describes the approach to constructing a speaker weight matrix using the multilinear decomposition of the acoustic model learned using the development data set. Tucker decomposition is employed, which, due to the merits of extracting shared information across the modes of the input tensor, yields the speaker subspace along with subspaces for other modes, and a grouping procedure is described to come up with a weight matrix for adaptation purpose [63]. Further, an adaptation procedure is discussed, which helps adapt the weight matrix to a new speaker from the enrollment data.

3.4.1 Speaker space construction

Firstly, we train a Universal Background Model (UBM) using the entire training set of speakers. This gives us a mean matrix of dimensions $S \times D$, where S is the total number of mixtures and D is the dimension of the mean vectors. Furthermore, mean matrices for individual speakers from the training set is calculated using Gaussian Mixture Modelling (GMM), which results in N number of mean matrices. To construct a speaker space based on Tucker Decomposition, the mean for the mixtures corresponding to each of the total N numbers of training speakers is collected. The mixture means, which are in the form of $S \times D$ matrix, are concatenated along the third dimension to form a tensor of dimensions $S \times D \times N$ denoted by $\mathcal{Y} \in \mathbf{R}^{S \times D \times N}$.

Now, the training tensor is decomposed using Tucker Decomposition into three subspaces corresponding to each mode and a core tensor.

$$\mathcal{Y} = \mathcal{X}^{S\times D\times N} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)} \tag{3.1}$$

Here, $\mathbf{U}^{(1)} \in \mathbb{R}^{S\times S}, \mathbf{U}^{(2)} \in \mathbb{R}^{D\times D}$ and $\mathbf{U}^{(3)} \in \mathbb{R}^{N\times N}$corresponds to the subspaces of mode-1(mixture), mode-2(dimensions) and mode-3(speaker) respectively and the core tensor provides for the coefficients of interaction between the three subspaces. The subspaces are considered to contain the variabilities pertaining to the information present in the modes. As such, the speaker subspace, along the third dimension of the tensor, exhibits speaker variabilities captured. The model for the speakers can be represented in tensor form as

$$\mu_n = \mathcal{Y}(:,:,n) = \mathcal{X} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}(n,:) \tag{3.2}$$

The μ_n thus obtained is a matrix of dimension $S \times D$ corresponding to the nth speaker. Hence, we have achieved a tensorial representation for each speaker model. Regrouping the components in the above equation, we get

$$\mu_n = \mathbf{U}^{(1)}.\{\mathcal{X} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}(n,:)\}^T \tag{3.3}$$

where '.' represents matrix product. Now, we merge $\mathcal{X}, \mathbf{U}^{(2)}$ and $\mathbf{U}^{(3)}$ into a single matrix. This matrix is the speaker weight matrix, denoted by $\mathbf{W}$ [71].

$$\mu_n = \mathbf{U}^{(1)}.\mathbf{W}_n^T \tag{3.4}$$

Thus it can be observed that the nth speaker is expressed as a weighted sum of the mixture subspace $\mathbf{U}^{(1)}$ and $\mathbf{W} \in \mathbb{R}^{D\times S}$.

Thus, in the enrollment step, for a new speaker, we need to adapt only the speaker weight matrix $\mathbf{W}$.

$$\mu_{new} = \mathbf{U}^{(1)}.\mathbf{W}_{new}^T$$

3.4.2 Enrollment speaker adaptation

Here, a speaker adaptation equation is derived based on the minimization of the probabilistic weighted distance of the test feature vectors from the mean of the mixture components. Given an utterance from a new enrollment speaker, feature vectors are calculated. As such, given adaptation

Table 3.1: Performance Evaluation measured in terms of Identification Accuracy and compared with baseline methods

Dataset	Enrollment(secs)	Testing(secs)	Accuracy (%)	
			MAP adaptation	Tensor adaptation
TIMIT	20	10	93.04	92.86
NIST SRE 2003	20	10	82.30	82.87
NIST SRE 2003	80	20	91.57	91.29

feature vectors $O = [o_1, o_2, \cdots o_t \cdots o_T]$, where each o_t is of dimensionality $\mathbf{R}^{D\times 1}$, the speaker weight is calculated by the minimization of the following error function [71]

$$Error = \sum_s \sum_t \gamma_s(t)|o_t - (\mathbf{U}^{(1)}(s,:).\mathbf{W}^T)^T|^2 \tag{3.5}$$

Here, the probabilistic weights i.e. $\gamma_s(t)$ is the posterior probability of the feature vector belonging to the mixture 's' at the instant 't' given the adaptation data 'O', and is calculated using the speaker-independent model i.e. the UBM trained using the background speaker data. Also, $\mathbf{U}^{(1)}(s,:)$ is the s-th row from the mixture subspace corresponding to the sth mixture component. As such, $(\mathbf{U}^{(1)}(s,:).\mathbf{W}^T)$ denotes the s-th mixture means.

Minimization of the error function requires the calculation of $\frac{\partial Error}{\partial \mathbf{W}} = 0$, which yields

$$\hat{\mathbf{W}} = \{\sum_s \sum_t \gamma_s(t) o_t.\mathbf{U}^{(1)}(s,:)\}$$
$$.\{\sum_s \sum_t \gamma_s(t)\mathbf{U}^{(1)}(s,:)^T.\mathbf{U}^{(1)}(s,:)\}^{-1} \tag{3.6}$$

In many practical cases, the training models are non-centered. As such, the adaptation equation is achieved by replacing (o_t with $o_t - \bar{\mu}(s,:)$), where $\bar{\mu}$ is the average over all the mixture mean matrices over all the speakers, i.e.

$$\bar{\mu} = \frac{1}{N}\sum_n \mu_n \tag{3.7}$$

thus the equation for the adapted models, in non centered case, becomes

$$\mu_{new} = \mathbf{U}^{(1)}.\mathbf{W}_{new}^T + \bar{\mu} \tag{3.8}$$

The advantage of such an adaptation over other adaptation methods such as EM algorithm etc., is that the requirement for adaptation data is relatively less. A good speaker model can be generated using only a few seconds of adaptation data. Moreover, if we use the truncated version of the speaker

space, fewer parameters need to be computed. Also another benefit is that the background model and speaker space are constructed offline. So, once the speaker subspace is learned, it can be adapted to any speaker population.

3.5 Experimental Analysis

The experiments are conducted using TIMIT and NIST SRE 2003 Datasets. While TIMIT is used for background model creation and subspace development, speaker adaptation is shown using the enrollment subset from both TIMIT and the NIST 2003 dataset. The metric used for performance evaluation is *Identification Accuracy*, that measures the number of correct predictions made by a model in relation to the total number of predictions made. We calculate it by dividing the number of correct predictions by the total number of predictions.

3.5.1 Datasets and Prepossessing

The TIMIT database is a high-quality, phonetically rich, microphone-recorded dataset often used for speech recognition experiments [72]. It is further divided into different dialect regions, and each speaker's data consists of around ten utterances of around three seconds duration each. The NIST SRE 2003 database consists of training and testing data from 356 speakers, recorded under telephone channel conditions [73]. The training and testing utterances are recorded in different sessions. The prepossessing stage consists of converting the speech utterances into 39-dimensional feature vectors consisting of 13-dimensional static mel coefficients and their derivatives.

3.5.2 Evaluation procedure

A Universal Background Model(UBM) is constructed consisting of 70 speakers from the TIMIT dataset, using an equal proportion of male and female speakers. The GMM size is fixed as 16. The baseline system used here is GMM-UBM using MAP adaptation on the same amount of data as used for the tensor case. For subspace creation, using 38 speaker subsets from TIMIT as development data, we learn the models for the speakers using MAP adaptation. Thus we get the mixture means corresponding to each of the 38 speakers having dimensions 16×39.

A tensor is constructed by keeping the number of speakers in the third mode while the first and second mode corresponds to mixture and dimensions, respectively. Thus we get the acoustic tensor model of dimensions $16 \times 39 \times 38$. Tucker decomposition of the acoustic tensor yields three subspaces

corresponding to the three modes, i.e., mixture subspace, dimension subspace, and speaker subspace of dimensions 16×16,39×39 and 38×38 respectively.

Enrollment and testing are done using TIMIT and NIST 2003. The log-likelihood is used to calculate the scores for the test feature vectors once the adapted means for enrollment speakers are obtained. We have considered two cases in adaptation. First, equal adaptation data and equal test data are used for both TIMIT and NIST 2003 speakers. Here, the adaptation is appreciable in the case of TIMIT speakers compared to NIST 2003 speakers. Secondly, adaptation and test data are increased for NIST 2003 speakers. In this case, the performance is comparable for both the speaker population. Also, it shows the channel adapting capability of the tensor representation of speaker space if adaptation data is sufficient. The tensor-based adaptation process performs comparably to the baseline GMM-UBM using MAP adaptation.

3.6 Conclusions

In this chapter, a tensor approach is described for the construction of speaker space. Tucker decomposition is used for the construction of spaces along the modes of the tensor, thereby preserving the distinction between mixture components and dimensions of mean vectors, which is lost in supervector-based methods. The subspaces consist of the basis along each mode unfolding. Regrouping the factor matrices provides us with a speaker weight to be updated during the adaptation of the new speaker. Experiments conducted on the TIMIT and NIST 2003 datasets show that the enrollment speaker adaptation yields appreciable results even in channel mismatch conditions if sufficient adaptation data is available. Furthermore, this method can be extended using higher order tensors with higher order modes consisting of noise, channel, session, etc., and a joint adaptation framework can be explored. Also, the tucker decomposition with other types of constraints, such as non-negativity, sparsity, etc., can be explored in learning a more efficient and robust subspace.

4

Tensor Factorization Based Neural Network Architectures for SER

Objective

In an attempt to make Human-Computer Interactions more natural, we propose the use of Tensor Factorized Neural Networks (TFNN) and Attention Gated Tensor Factorized Neural Network (AG-TFNN) for Speech Emotion Recognition (SER) task. Standard speech representations such as 2D and 3D Mel-Spectrogram and Temporal Modulation Spectrogram are explored to investigate the emotion salient information capturing effectiveness of the Tensor Factorization-based architectures. The hidden layers are explained as Deep Tensor Factorization based on the Tucker Decomposition but with a unified discriminative objective function to learn the factor matrices in a discriminative sense. The core tensor produced in each hidden layer is the feature associated with that factorization layer. Mel Spectrograms are naturally in 2D tensor form, and thus TFNN and AG-TFNN become an appropriate choice over baselines such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) by providing reduced parameters to be learned and simple architecture. Experiments conducted on standard emotional speech datasets- Emo-DB and IEMOCAP show that TFNN and AG-TFNN surpass the state-of-the-art CNN+LSTM combination with fewer parameters.

4.1 Introduction and Related work

The use of tensor factorization in feature learning and parameter reduction has caught the attention of researchers for a while. Many works have proposed using tensors in combination with deep learning architectures to reduce the trainable parameters between the layers. In [74], a tensor version of the Restricted Boltzmann Machine (RBM) is proposed to model the three-way interaction between the pixels of an image. A similar approach has been proposed in [75] to model the pixel means and covariances. Also, the work in [76] generalize RBMs to capture the multiplicative interaction between data modes and the latent variables and presents a successful application in handwritten digit recognition, face recognition, and EEG-based alcoholic diagnosis.

The idea of tensor factorization in hidden layers was further explored in [77]. This work extended the idea of a conventional DNN by replacing one or more layers with a double-projection (DP) layer. The technique is to project the input vectors into two nonlinear subspaces, followed by a tensor layer, in which two subspace projections interact with each other and jointly predict the next layer in the deep architecture. Application of this DTNN on large vocabulary speech recognition using switch-board dataset showed improvement over the baseline. Building on this idea, [78] proposed Multilinear

Principal Component Analysis Network (MPCAnet), which is a tensor extension of PCANet [79]. The aim was to extract high-level features from multi-dimensional images otherwise lost in traditional methods.

Recent work in [80] proposed a Tensor Factorized Neural Network (TFNN), which tightly integrated the feature extraction and classification of tensor data under a unified discriminative objective function. TFNN helps reduce the parameter size to a large extent and can learn discriminative weights for each hidden layer. TFNN is naturally suited for tasks with 2D or higher-order tensor inputs and, as such, becomes an apt choice for SER, as explained below.

4.2 Motivation for using TFNN in SER

Spectrograms and similar representations like Mel-spectrograms, MFCCs, etc., are in 2D form [27] and usually, the emotional label is available for the entire speech utterance rather than frame/phoneme based. Moreover, recent works have explored 3D -Mel Spectrograms as an input feature to a 3D CNN network [27]. Vectorization of such 2D and 3D representations causes the problem of high dimensionality, increased number of parameters in each layer, and the spatial relation of the pixel values getting lost. This property becomes essential when dealing with spectrograms as they represent the energy present in the different frequency components over time instants. This, in turn, can be leveraged for SER since it has been shown in the literature that emotions and frequency variations are highly correlated [81]. Vectorization breaks this correlation among the energy components at progressing time instants, which is of the utmost importance when characterizing emotional information in speech utterances. Hence an algorithm to deal with the 2D and 3D structure of speech representations without breaking the correlation among the pixels is required. In deep learning, Convolutional Neural Network (CNN) is one such architecture developed for dealing with 2D and 3D structures as input with the help of 2D and 3D kernels. However, it has disadvantages such as a vast number of parameters with increasing layers, dimensions, and kernels, losing information because of pooling layers, and Vectorization of deep features for classification using fully connected layers. As the number of parameters increases, the data requirement also increases proportionately. TFNN architecture is blessed with the merits of keeping the 2D and 3D forms intact. The number of parameters is appreciably low even with an increase in the number of hidden factorization layers and the dimension of the input tensors.

Also, deep learning architectures are often attributed as “Black Boxes” as proper visualization

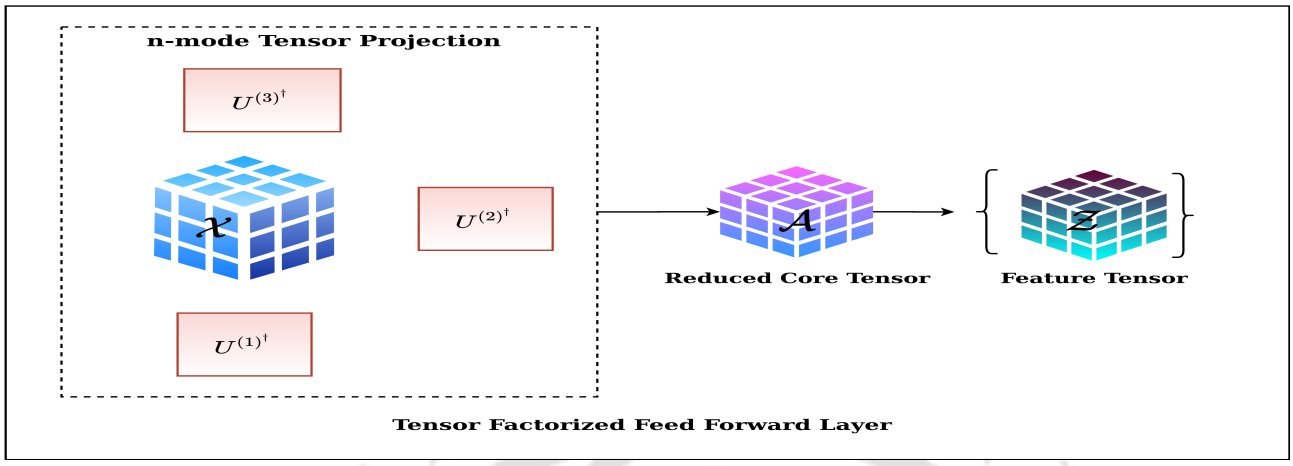


Figure 4.1: A Tensor feed-forward block. Input tensor $\mathcal{X} \in \mathbf{R}^{I_1 \times I_2 \times I_3}$ is factorized by mode-n factor matrices $U^{(n)}, n = 1, 2, 3$ to get reduced the core tensor $\mathcal{A}$, which is passed through an element-wise non-linear activation function to get the feature tensor $\mathcal{Z}$

and explanation of hidden layers are pretty obscure. However, the predictive power of deep learning algorithms triumphs over the demerit of explanatory powers. The TFNN has the benefit of both explanatory power, and predictive power [80]. The hidden layers can be visualized as repeated tensor factorizations as per Tucker Decomposition, except that the factor matrices learned are discriminative. This is achieved by virtue of a discriminative objective function rather than a least-square objective function, as is used for tensor reconstruction in traditional Tucker decomposition. Thus the richness of tensor algebra could serve as an explanation for the deep feature tensors in the case of TFNNs.

4.3 Tensor Neural Network description

Tensor Factorized Neural Network was introduced in [80] with the motivation of preserving multi-way information through layerwise tensor factorizations. The amalgamation of Tensor factorization and a Neural network is driven by the idea that n-way tensor factorization of an n-Dimensional Tensor helps in capturing the common factors present across the different ways, which can be further modeled in a neural network framework to produce posterior outputs by minimizing a cross-entropy error. This section describes in detail the feed-forward and the attention layer for an n-Dimensional Tensor.

4.3.1 Tensor Feed Forward Layer

Tensor Feed Forward block, as shown in Figure 4.1, utilizes the notion of Tucker Decomposition to obtain a core Tensor, which contains the coefficients of interaction among the n-way factors. Given

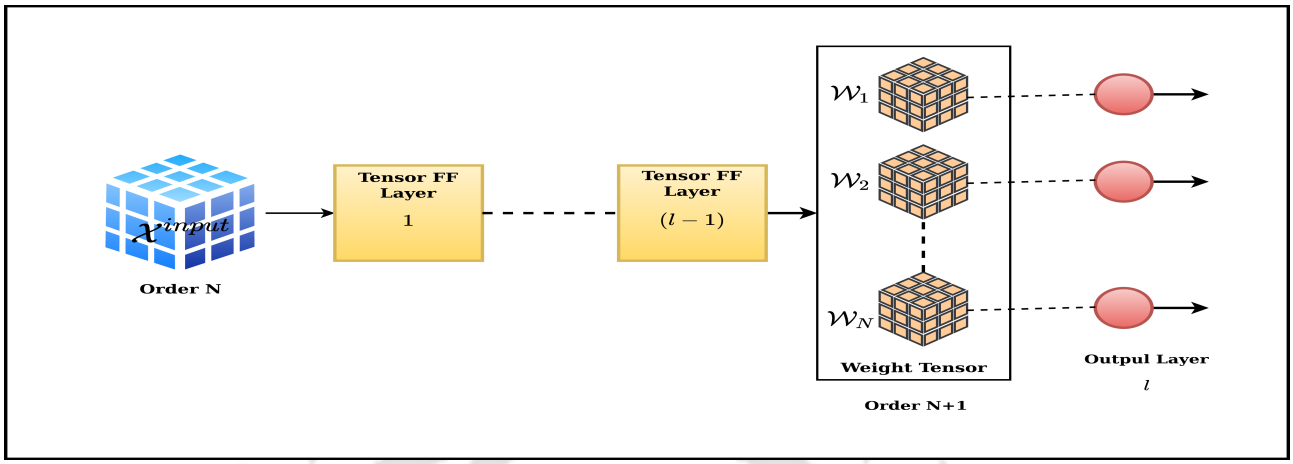


Figure 4.2: The complete Tensor Factorized Neural Network architecture. Input Tensor $\mathcal{X}$ is feed-forward through tensor-factorization blocks to obtain a deep feature tensor. The weight tensor $\mathcal{W}$ is projected upon the feature tensor, and the output is passed through a softmax function to obtain class probabilities.

an n-Dimensional input Tensor $\mathcal{X} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_N}$, it is converted to core-tensor $\mathcal{A} \in \mathcal{R}^{J_1 \times J_2 \times \cdots \times J_N}$, where $J_n \leq I_n$ by utilizing the mode-n product -

$$\mathcal{A} = \mathcal{X} \times_1 \mathbf{U}^{(1)^\dagger} \times_2 \mathbf{U}^{(2)^\dagger} \times_3 \cdots \times_N \mathbf{U}^{(N)^\dagger} \tag{4.1}$$

Here, $\mathbf{U}^{(n)} \in \mathcal{R}^{I_n \times J_n}$ are the mode-n factor matrices (not necessarily orthogonal) along the n-ways of the input tensor. The factor matrices provide the connections from the input tensor to the core tensor along the n-ways. The obtained core tensor is passed through an activation function to obtain the hidden layer output as follows:

$$\mathcal{Z}^l = func(\mathcal{A}) \tag{4.2}$$

where $func$ is a non-linear activation function such as $RELU$ (Rectified Linear Activation Unit), $tanh$, $sigmoid$ etc and $\mathcal{Z}^l$ is the deep feature tensor for the l-th layer. In our work, we have used $RELU$ activation function, which is applied element-wise to the entries of the core tensor to obtain the activated hidden layer output.

The Tensor Factorized Neural Network is built up stacking the Feed Forward Blocks, as shown in Figure 4.2. An input tensor $\mathcal{X}$ is passed through feed-forward blocks to obtain the activated deep feature tensors $\{\mathcal{Z}^1, \mathcal{Z}^2, \cdots \mathcal{Z}^{l-1}\}$, where l-th layer is the last layer of the network, usually the softmax layer. The last layer provides the $\mathcal{C}$ posterior classification outputs $\mathbf{y} = \{y_c\}$ by using a softmax function, where $c \in 1, \cdots, C$ and C being the number of classes. The vector of class posteriors is

obtained by projecting the N-Dimensional feature tensors onto a $(N+1)$-Dimensional weight tensor $\mathcal{W} \in \mathcal{R}^{k_1 \times k_2 \times \cdots \times k_N \times C}$. The l-th layer, which is the last layer, can be mathematically described as -

$$\mathbf{a}_c^l = \langle \mathcal{W}_{:,:,\cdots,c}, \mathcal{Z}^{l-1} \rangle \tag{4.3}$$

where $\mathbf{a}_c^l$ is an entry of the vector $\mathbf{a} \in \mathcal{R}^{C \times 1}$ and is obtained by the inner-product (for definition see [44]) of the c-th N-D sub-Tensor of the $(N+1)$-D weight tensor $\mathcal{W}$. $\mathcal{Z}^{l-1}$ is the activated feature tensor from the $(l-1)$-th layer and have the dimensions $\mathcal{R}^{k_1 \times k_2 \times \cdots \times k_N}$. The class probabilities corresponding to the C-output classes are obtained by passing the vector $\mathbf{a}$ through a softmax function as follows

$$\mathbf{y} = softmax(\mathbf{a}^l) = \frac{\exp(\mathbf{a}^l)}{\sum_{c=1}^{C} \exp(a_c^l)} \tag{4.4}$$

The factor matrices $\{\mathbf{U}^1, \mathbf{U}^2, \cdots, \mathbf{U}^N\}$ for each of the hidden layer and the weight tensor $\mathcal{W}$ of the last layer accounts for the parameters of the TFNN to be trained through error backpropagation. Given a set of training tensors and label pairs $\{\mathcal{X}_t, \mathbf{r}_t\}$, the error function to be minimized is the cross-entropy error, which is given by -

$$\mathbf{E} = \sum_{t=1}^{T} \mathbf{E}_t = -\sum_{t=1}^{T} \sum_{c=1}^{C} r_{tc} \ln y_{tc} \tag{4.5}$$

where, $\mathbf{y}_t = \{y_{tc}\}$ is the output vector from the softmax layer, pertaining to the C-class probabilities, $\mathbf{r}_t = \{r_{tc}\}$ is the target label vector of the t-th training sample in one-hot encoding.

4.3.2 Tensor Attention Layer

The Tensor Attention Layer focuses on the emotion-relevant parts of the input spectrogram by learning attentive weights and thus producing emotionally-discriminative processed input for the subsequent Tensor Feed-Forward Layers to learn higher level features. In our work, we have used the Attention layer at the beginning of the network to give the network more emotion-focused input, which helps to model the emotions more effectively and thus reduces confusion.

Specifically, as shown in Figure 4.3, with the input tensor $\mathcal{X} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_N}$, we compute the attentive core tensor $\mathcal{A} \in \mathcal{R}^{J_1 \times J_2 \times \cdots \times J_N}$, where $J_n \leq I_n$ by projecting on the factor matrices $\mathbf{U}^{(n)} \in \mathcal{R}^{I_n \times J_n}$ which serves as the attentive weights and is learnt in the back-propagation framework. The core tensor $\mathcal{A}$ is passed through a softmax function to get the importance of each entry of the tensor $\mathcal{X}$. The attentive tensor, which serves as the input for the subsequent layers of TFNN, is obtained

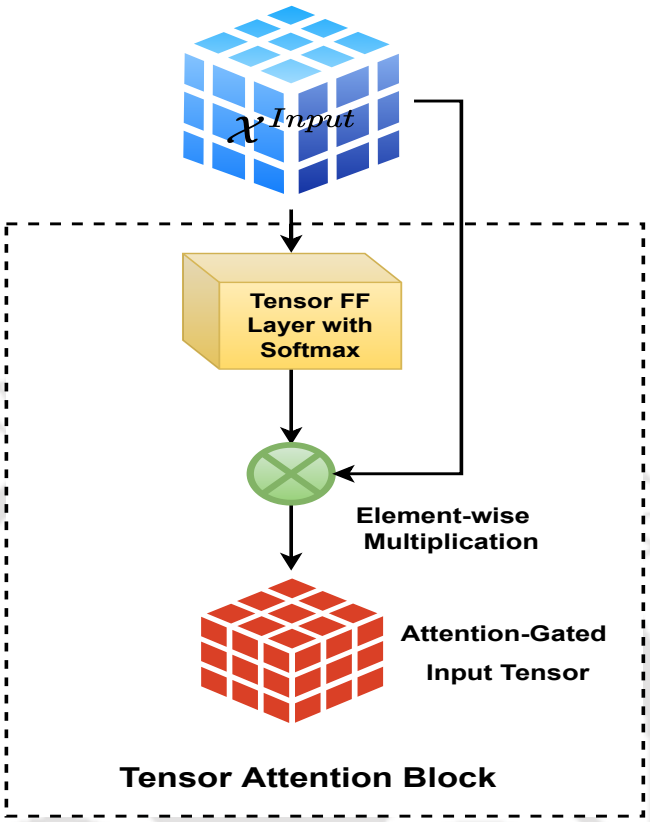


Figure 4.3: Tensor Attention Mechanism.

by multiplying the input tensor $\mathcal{X}$ with the importance weights. Mathematically, using 4.1, the core tensor is calculated as -

$$\mathcal{A} = \mathcal{X} \times_1 \mathbf{U}^{(1)^\dagger} \times_2 \mathbf{U}^{(2)^\dagger} \times_3 \cdots \times_N \mathbf{U}^{(N)^\dagger} \tag{4.6}$$

The core tensor $\mathcal{A}$ is passed through a softmax function to get the importance weights of each element of the core tensor -

$$\mathcal{A}_{imp} = softmax(\mathcal{A}) \tag{4.7}$$

Finally, to get the attention weighted input for the subsequent Tensor FF Layers, the original input tensor $\mathcal{X}$ is elementwise multiplied with the importance weights generated using 4.7 -

$$\mathcal{X}_{attentive} = \mathcal{X} \odot \mathcal{A}_{imp} \tag{4.8}$$

Where $\odot$ represents the element-wise multiplication of two matrices of equal size. The obtained attentive input $\mathcal{X}_{attentive}$ serves as the input to further Tensor FF layers.

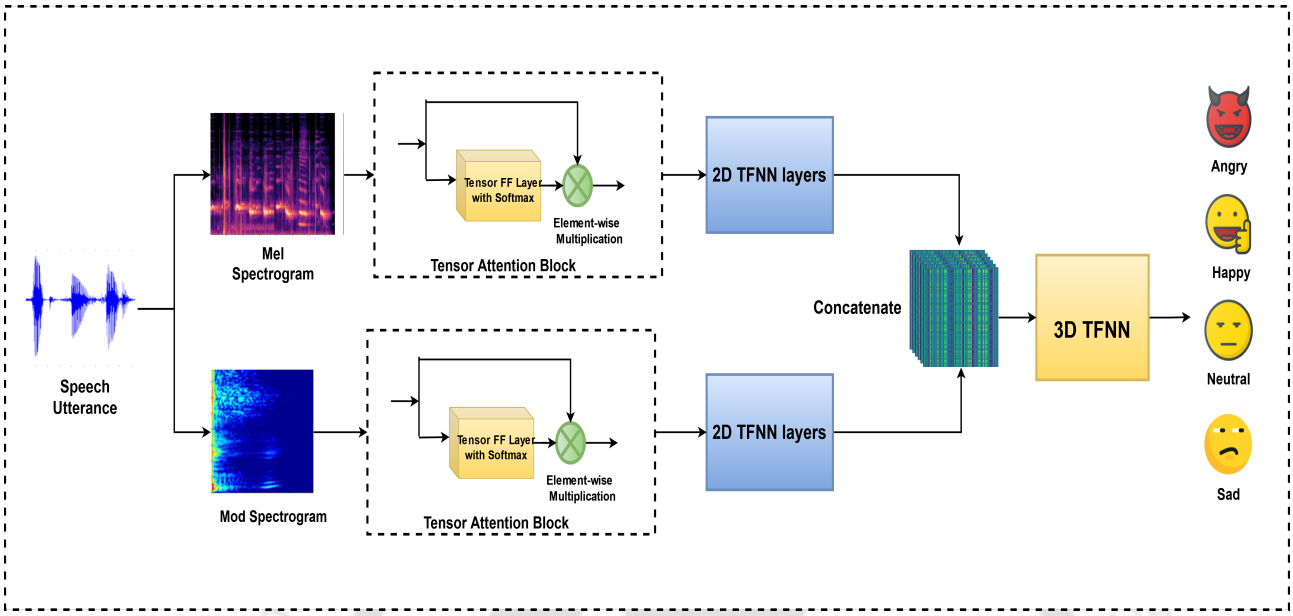


Figure 4.4: The proposed parallel AG-TFNN network for SER using complementary features from mel-spectrogram and modulation-spectrogram.

4.4 Parallel AG-TFNN

In order to leverage complementary information from speech representations such as Mel-Spectrogram and Modulation Spectrogram, we propose a parallel AG-TFNN architecture. Mel-Spectrograms depict the frame-wise energy content over the range of acoustic frequency. On the other hand, Modulation Spectrogram is known to capture the temporal frequency modulation in a signal. Both contain different information, which can be utilized in a unified architecture for SER. The parallel AG-TFNN architecture is shown in Figure 4.4. It consists of two parallel 2D AG-TFNN networks responsible for extracting high-level representations from the two input 2D tensors. The feature tensors are projected using factor matrices to be of the same size. The high-level feature tensors are concatenated to form a 3D tensor and are fed to a 3D Tensor FF layer. This concatenation ensures that shared common features are extracted further by virtue of 3D Tensor FF Layers. A 3D core tensor, the new feature representation, is obtained and passed on to the softmax layer for generating class probabilities.

4.5 Baseline Architecture Description

The baseline used in our work to compare the classification performance of the Tensor Factorized Neural Network(TFNN) is the combined 2D CNN+LSTM architecture as described in [26]. The

2D CNN+LSTM architecture is constructed using four Local Feature Learning Blocks (LFLBs), one LSTM layer, and one Fully-connected layer with softmax activation to yield the class probabilities. Each of the four LFLBs comprises one each- convolutional layer, batch normalization layer, activation layer, and max-pooling layer. The activation function selected is an exponential linear unit (ELU). The LFLBs are responsible for extracting the local spatial information from 2D speech representations such as spectrograms. The convolutional layers have the advantage of local spatial connectivity and shared weights, which enables them to perform kernel learning. The batch normalization layer is applied to the pre-activations in order to normalize the activations of the convolutional layers by maintaining the mean to 0 and standard deviation close to 1. This improves stability and speeds up the training process [82]. The activation is chosen as ELU as it has negative values also compared to $RELU$, which only has positive ones. Thus it pushes the mean of the activations towards zero, thereby speeding up the learning process [83]. Max pooling is used to reduce the size of the feature maps produced, consequently scaling down the number of trainable parameters in the subsequent layers.

Global feature learning is performed by LSTM to capture the long-term temporal dependencies in an utterance. The local learned features from the stacked LFLBs are passed on to the LSTM layer to learn the long-term contextual dependencies. The size of convolutional kernels is taken as 3×3, and a stride of 1×1 is used. The number of convolutional kernels in the four LFLBs are 64, 64, 128, and 128, respectively. Max-pooling with kernel size and stride of 2×2 is used in the first LFLB and 4×4 in the other blocks. The output from the LSTM layer contains both the local correlations and the global contextual dependencies, which are passed on to the fully connected (FC) layer for classification. The softmax function, used as the activation in the FC layer, yields the class probabilities of the different emotional classes.

4.6 Dataset Description

The Tensor Factorized Neural Network(TFNN) is applied to the task of Speech Emotion Recognition (SER). Two datasets are chosen to demonstrate the effectiveness of the Tensor Factorization-based architectures in capturing the emotion salient information in the speech utterances. The Datasets are chosen based on popularity among the researchers, the degree of naturalness of the elicited emotions and size of the dataset, and the distribution of utterances across the emotion categories. Standard speech representation such as 2D and 3D mel-spectrogram is used as an input feature to the TFNN,

AG-TFNN, and CNN+LSTM, and modulation spectrogram is used additionally as an input to the parallel AGTFNN and Weighted Accuracy(WA) as well as Unweighted Accuracy(UA) of the utterances are reported as a classification metric.

4.6.1 Emo-DB

The Berlin Emotional database, i.e., the Emo-DB database, is one of the most widely used databases in Speech Emotion Recognition studies. It comprised seven emotion categories- Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral and was recorded in 2005 by ten German actors - five males and five females, in a single session. Ten sentences - five long and five short sentences, were chosen to contain the emotion salient and phonetic information balanced. The signals were recorded at a sampling rate of 48 kHz and later downsampled to 16 kHz. The utterances are of duration 1 to 4 secs with an average duration of 3 secs. The dataset consists of a total of 535 utterances spread over seven emotion categories.

4.6.2 IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) is an English emotional dataset comprising both motion data and audio clips recorded by five pairs of actors (male-female). It contains approximately 12 hours of audiovisual data consisting of video, speech, motion capture of face, text transcriptions, etc. The dataset is recorded in two scenarios- scripted and improvised, specifically selected to elicit emotional expressions. Multiple annotators annotate the dataset into categorical labels such as Angry, Excited, Frustrated, Happiness, Neutral, Sadness, and Surprise and dimensional labels such as valence, activation, and dominance. The total utterances from both the scripted and improvised sessions come up to around 5530 for the four chosen emotion categories (Angry, Happy, Neutral, and Sadness).

4.7 Feature extraction and Tensor Formation

We investigated the proposed AG-TFNN architecture for both 2D and 3D input tensors. The different speech representations utilized for tensor formation are discussed in brief below:

4.7.1 2D Tensors - log mel Spectrograms and Modulation Spectrograms

Two different speech representations are investigated to be used as 2D input for the TFNN and AG-TFNN architecture- Log Mel-Spectrogram (will be referred to as Mel-spectrogram) and Modulation

Spectrogram(will be referred as Mod-spectrogram). Spectrograms provide a 2D representation in the form of a time-frequency matrix displaying the energy of different frequency components at varying time instants. In our work, spectrograms are computed from the speech signal by first segmenting and windowing the speech signal using a Hanning window. The window size is taken as 2,048 samples with an overlap of 512 samples, and Short-Time Fourier Transform (STFT) is calculated for windowed segments. As such magnitude spectrogram is obtained by calculating the energy of the frequency components obtained by STFT.

Mel-spectrograms have seen a recent surge in usage in SER tasks using Deep learning [27] [84]. The motivation behind using Log-Mel spectrograms over Magnitude spectrograms is that the Mel-scale is believed to simulate the perception of human ears better by emphasizing the lower frequency region more than the higher frequency region [85]. This is achieved by having more number of mel filters in the lower frequency region. For calculating the Mel-spectrogram, the magnitude spectrogram is first obtained and then mapped to a mel scale to obtain the filter bank energies passed through a log operator. The number of mel frequency bands used in our work is 128.

Moreover, since the emotion salient information is spread over the entire utterance duration, the speech signal's temporal modulation is of significance when dealing with SER tasks. This motivated the use of temporal modulation spectrograms from speech utterances. The evaluation of the modulation spectrogram consists of two stages. In the first stage, magnitude spectrograms are calculated from the speech utterances by windowing the signal and finding the STFT. In the second stage, each row of the magnitude spectrogram, which represents a frequency component over the entire time instant, is processed using wavelet transform or STFT to retrieve the temporal modulation content.

In pre-processing, the speech utterances are either zero-padded or chopped into equal length segments of size 50,000 samples (3.12 seconds approx.) as the input to the TFNN requires equal-sized tensors. The equal-sized feature representations obtained are further normalized to have values between 0 and 1.

4.7.2 3D Tensors - 3D log mel spectrograms

For the construction of 3D mel-spectrograms from speech utterances to be used as input to 3D AG-TFNN, the deltas and double-deltas are computed and stacked along the third axis to form a 3D tensor. In this process, the speech signal is split into a number of short frames with a hamming window of frame size 2,048 samples and a frame shift of 512 samples. The power spectrum for each frame

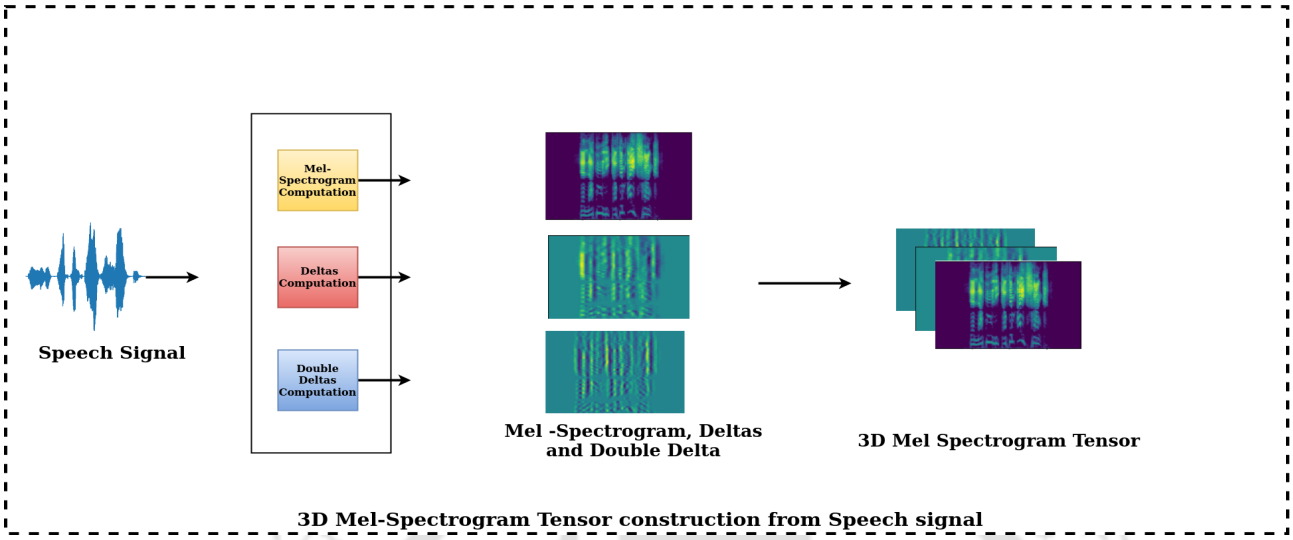


Figure 4.5: 3-D Tensor construction from a speech utterance. 3D Log-Mel Spectrogram tensor is constructed by stacking Mel-spectrogram, deltas, and double-deltas

is calculated using STFT and passed through a mel-filter i bank to obtain filter output P_i. Taking the logarithm of the mel-filterbank outputs yields the log-mels m_i. Then the deltas are computed by taking the time derivative of the log-mels, and the double-deltas are computed by taking the time derivative of the deltas. After stacking the log-mel spectrogram, the deltas, and double deltas, we obtain a Mel spectrogram tensor $\mathcal{X} \in \mathcal{R}^{filter \times time \times channel}$, where $filter$ denotes the number of mel frequency bands, $time$ denotes the frame length, and $channel$ denotes the number of feature channels in the tensor, $c = 3$ in our case. Figure 4.5 shows the formation of 3D mel spectrogram tensors for an utterance belonging to Angry emotion of the Emo-DB dataset.

4.7.3 Experimental setup

The speech representations such as spectrograms, Mel-Spectrograms, and Modulation Spectrograms are second-order tensors or popularly known as matrices, and 3D Log-Mel Spectrograms are third-order tensors. The values of the feature matrices are normalized to the range between 0 and 1. The factor matrices, which serve as the n-way weights in the hidden layers, are initialized randomly from a uniform distribution with zero mean and unit variance. Cross-entropy loss is used to calculate the loss for error backpropagation, which ensures that the factor matrices learned in the hidden layers are discriminative.

The speech representations such as 2D and 3D log-mel spectrograms are calculated using the

python library Librosa [86] and Modulation Spectrogram using AM Analysis Toolbox[1]. The TFNN and the baseline CNN+LSTM are implemented using the python deep learning library Keras with the backend TensorFlow. Bayesian optimization for selecting hyper-parameter values were performed using the library scikit-optimize[2].

The experiments are divided into two categories- Speaker Dependent(SD) and Speaker Independent (SI). For both datasets, each of the two scenarios is evaluated. In Speaker-Dependent (SD) experiments, the dataset of four emotions, i.e., Angry, Happy, Neutral, and Sadness, is randomly split into two parts with 80 % for training and 20 % for testing. The splits are made using five random seeds, and the mean and standard deviation of the accuracies across the five-fold is reported. In the Speaker-Independent(SI) experiments, since ten actors record both Emo-DB and IEMOCAP, we have taken data from eight speakers as training with the remaining two speakers for validation and testing. SI experiments are performed on ten sets of data for an input feature, and the mean and standard deviation over the ten speaker independent sets have been reported.

The number of Tensor FF Layers and the reduction in the size of tensor in each layer is optimized using Bayesian Optimization [87]. The range considered for the number of Tensor FF layers is $(0, 9)$, and for the reduction in the size of tensors in each layer is $(5, 20)$. The TFNN architecture consists of four Feed Forward layers and one softmax layer. The AG-TFNN architecture consists of one Attention layer, three Feed Forward Layers, and one softmax layer. The layer sizes and the number of parameters to be learned are summarised in Table 4.1 and Table 4.2. The inputs to the TFNN network are Mel-spectrogram of size 128×98 and Modulation spectrogram of size 161×156. For the 3D AG-TFNN case, the input is 3D log-mel spectrogram of size $128 \times 98 \times 3$.The Layer and parameter sizes are listed in Table 4.3.

4.7.4 Optimization and Model Selection

The TFNN as well as the CNN+LSTM architecture is optimized using Adam optimizer [88], [89]. Adam is selected because it combines the advantage of both the RMSprop and AdaGrad optimizers. Adam calculates individual adaptive learning rates for different parameters based on the estimates of the first and second moments of the gradients. The values for the parameters of Adam are $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with an initial learning rate of 0.001. Batch size for Emo-DB and IEMOCAP is taken

[1] https://github.com/MuSAELab/amplitude-modulation-analysis-module
[2] https://scikit-optimize.github.io/stable/

Table 4.1: The layer parameters of TFNN architecture with Log-Mel Spectrogram as Input. FF represents Feed Forward Layer as described in 4.3.1.

Sl.no	Layer Type	Output Shape	No.of Parameters
1.	Input Layer	128 x 98	0
2.	Tensor FF Layer	120 x 90	24,180
3.	Tensor FF Layer	110 x 80	20,400
4.	Tensor FF Layer	100 x 70	16,600
5.	Tensor FF Layer	90 x 60	13,200
6.	Softmax Layer	4	21,600

Table 4.2: The layer parameters of Attention-TFNN architecture with Log-Mel Spectrogram as Input. FF represents Feed Forward Layer as described in 4.3.1.

Sl.no	Layer Type	Output Shape	No.of Parameters
1.	Input Layer	128 x 98	0
2.	Attention Layer	128 x 98	25,988
3.	Tensor FF Layer	120 x 90	24,180
4.	Tensor FF Layer	110 x 80	20,400
5.	Tensor FF Layer	100 x 70	16,600
6.	Softmax Layer	4	28,000

as 15 and 25 respectively and is selected automatically using Bayesian Optimization [87]. Batch size significantly influences the generalization and convergence capability of the model. A small batch size increases model generalization to unseen data by ensuring model convergence to flat minimizers. On the other hand, a large batch size ensures model convergence to sharp minimizers resulting in the poor generalization of the models and degradation of the model capability in dealing with unseen data [90].

Model selection is employed to record the best-trained model with a superior predictive performance, which happens when the validation accuracy stops increasing during model training. However, training accuracy did not reach its maximum when the validation accuracy stopped increasing. As

Table 4.3: The layer parameters of 3D AG-TFNN architecture with 3D Log-Mel Spectrogram as Input. FF represents Feed Forward Layer as described in 4.3.1.

Sl.no	Layer Type	Output Shape	No.of.Parameters
1	Input Layer	128 x 98x3	0
2	Attention Layer	128 x 98 x 3	25,997
3	Tensor FF Layer (3D)	120 x 90 x3	24,210
4	Tensor FF Layer (3D)	110 x 80 x 3	20,500
5	Tensor FF Layer(3D)	100 x 70 x 3	16,630
6	Softmax Layer	4	84,000

Table 4.4: Weighted Accuracies (WA) and Unweighted Accuracies(UA) for Speaker Dependent(SD) experiments using CNN+LSTM and the proposed architectures TFNN, AG-TFNN, 3D AG-TFNN, and Parallel AG-TFNN on Emo-DB dataset.

Method	Speaker Dependent	
	WA	UA
CNN+LSTM	85.50 (3.58 %)	83.47(3.94%)
TFNN	86.18(3.30%)	85.48(3.12%)
AG-TFNN	88.53(3.28%)	86.97(4.57%)
3D AG-TFNN	**91.18(2.79%)**	**90.44(3.11%)**
Parallel AG-TFNN	83.53(4.69%)	81.57(6.82%)

such, overfitting occurs when validation accuracy decreases even when training accuracy increases, which is an undesirable situation. To counter this, early stopping of model training is employed, which can prevent overfitting and thus improve the model's predictive performance. The monitor for early stopping in our work is validation accuracy with patience of ten epochs. As such, the model trains for ten more epochs before the training stops when validation accuracy is not increasing. The best model is obtained when the validation accuracy is the highest, and that model is used for the performance analysis using the test utterances.

4.8 Results and Discussions

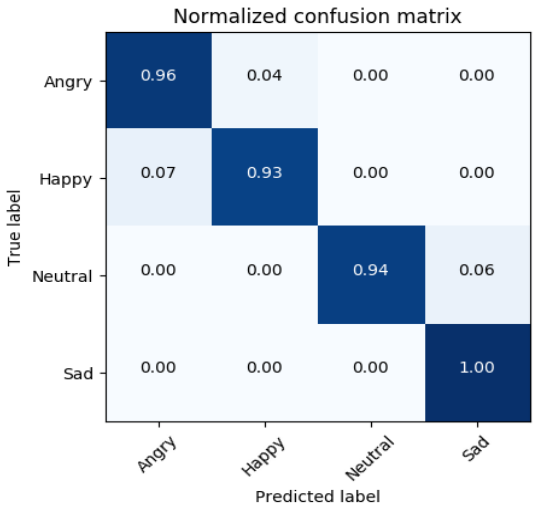
The task of SER is evaluated on EMO-DB and IEMOCAP using the TFNN architecture, 2D and 3D AG-TFNN architecture, and Parallel AG-TFNN architecture and is compared with the standard baseline method CNN+LSTM architecture. Contrary to the conventional deep learning techniques such as CNN and LSTMs, where the focus is more on the predictive power of the network rather than the explanatory power, in TFNN, each hidden layer can be visualized as a Tucker decomposition of an N-Dimensional tensor into subspaces comprising of the modes of the tensor and a reduced coefficient tensor called as a core tensor. The mode-n factor matrices or the subspaces are learned iteratively in a discriminative fashion rather than in the least square sense compared to traditional Tucker Decomposition. Both Speaker-dependent(SD) and Speaker-independent(SI) experiments are performed, and the mean and standard deviation of recognition accuracies are reported with the confusion matrix.

Table 4.5: Weighted Accuracies (WA) and Unweighted Accuracies(UA) for Speaker Independent(SI) experiments using CNN+LSTM and the proposed architectures TFNN, AG-TFNN, 3D AG-TFNN, and Parallel AG-TFNN on Emo-DB dataset.

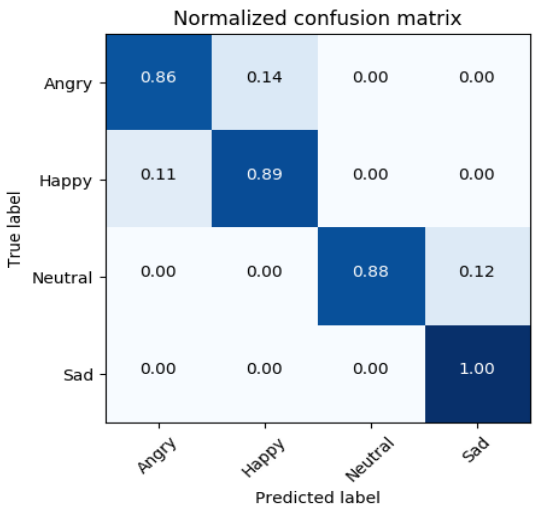
Method	Speaker Independent	
	WA	UA
CNN+LSTM	86.12(4.01%)	84.78(4.25%)
TFNN	83.54(7.97%)	81.22(8.56%)
AG-TFNN	81.86(6.65%)	81.30(8.11%)
3D AG-TFNN	**85.56(6.06%)**	**85.15(6.45%)**
Parallel AG-TFNN	81.65(5.66%)	81.14(5.72%)

4.8.1 Performance analysis using Emo-DB

The Emo-DB dataset consists of 271 train utterances and 68 test utterances, including four emotion classes (Angry, Happy, Neutral, and Sadness). The dataset is randomly split into train and test sets to remove any bias towards specific utterance sets. Table 4.4 and 4.5 shows the performance of the TFNN, 2D and 3D AG-TFNN, Parallel AG-TFNN, and CNN+LSTM architecture in capturing emotion salient information on Emo-DB using speech representations such as 2D and 3D Mel-spectrogram and Modulation Spectrogram for the Speaker-Dependent(SD) and Speaker Independent(SI) scenario respectively. The tables show that the TFNN and 2D and 3D AG-TFNN models can generalize well using the Mel Spectrogram features in the SD scenario outperforming the baseline CNN+LSTM. The best performance is obtained using 3D AG-TFNN architecture with 3D mel spectrogram as input surpassing the baseline by 5.68 percentage point in WA and 6.97 percentage point in UA. As evident from the confusion matrix in Figure 4.6a, there is very minimum confusion among the classes, with all four classes being noticeably learned by the model. However, the Parallel AG-TFNN architecture cannot surpass the baseline CNN+LSTM for the Emo-DB dataset. For the SI scenario, the performance of TFNN and AG-TFNN is comparable to the baseline CNN+LSTM, if not surpassing. The 3D AG-TFNN architecture surpasses the baseline by a margin of 0.37 percentage point in terms of UA. However, 3D AG-TFNN cannot surpass baseline CNN+LSTM in terms of WA. The confusion matrix reported in Table 4.6b shows that the maximum confusion occurs between the Happy and Angry classes, with 14% of the utterances belonging to the Angry emotion category being classified as Happy emotion.

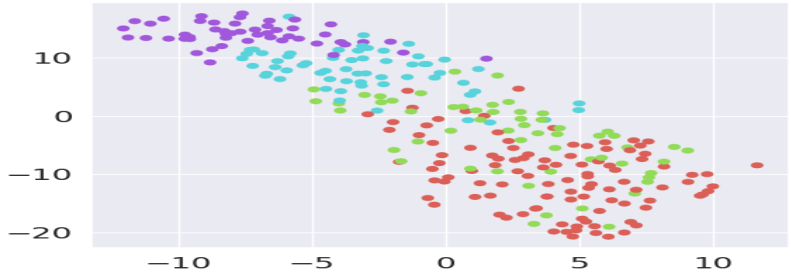


(a) Speaker Dependent

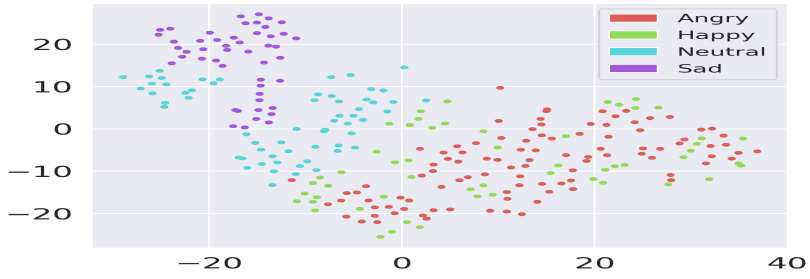


(b) Speaker Independent

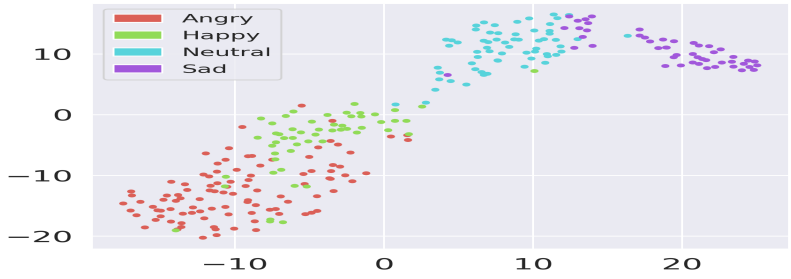
Figure 4.6: Normalized Confusion matrix for 3D AG-TFNN architecture using 3D Log Mel Spectrogram for (a) Speaker Dependent and (b) Speaker Independent Scenario on Emo-DB dataset



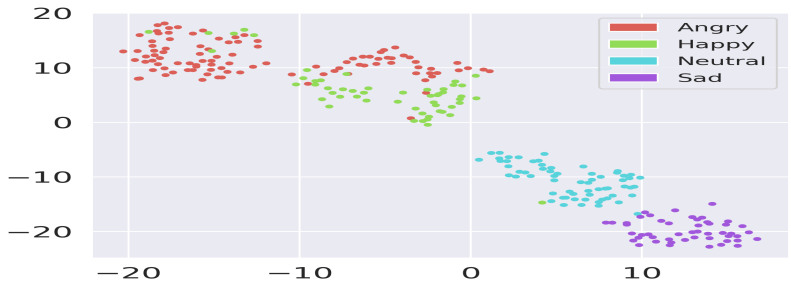
(a) Untrained Distribution of Emotions



(b) Embeddings from CNN+LSTM network



(c) Embeddings from 3D AG-TFNN network



(d) Embeddings from Parallel AG-TFNN network

Figure 4.7: T-sne scatter plot showing the distribution of the training set of Emo-DB using Mel-spectrogram as input for the four emotion classes. 4.7a displays the untrained distribution, 4.7b displays distribution after training by the CNN+LSTM architecture 4.7c displays the distribution after training by 3D AG-TFNN architecture, and 4.7d displays the distribution after training by Parallel AG-TFNN architecture.

Table 4.6: Weighted Accuracies (WA) and Unweighted Accuracies(UA) for Speaker Dependent(SD) experiments using CNN+LSTM and the proposed architectures TFNN, AG-TFNN, 3D AG-TFNN, and Parallel AG-TFNN on IEMOCAP dataset.

Method	Speaker Dependent	
	WA	UA
CNN+LSTM	53.52(2.23%)	54.18(1.25%)
TFNN	52.86(1.16%)	54.6(1.19%)
AG-TFNN	53.93(0.94%)	56.01(0.89%)
3D AG-TFNN	55.36(1.24%)	57.63(0.33%)
Parallel AG-TFNN	**56.95(1.08%)**	**59.39(1.39%)**

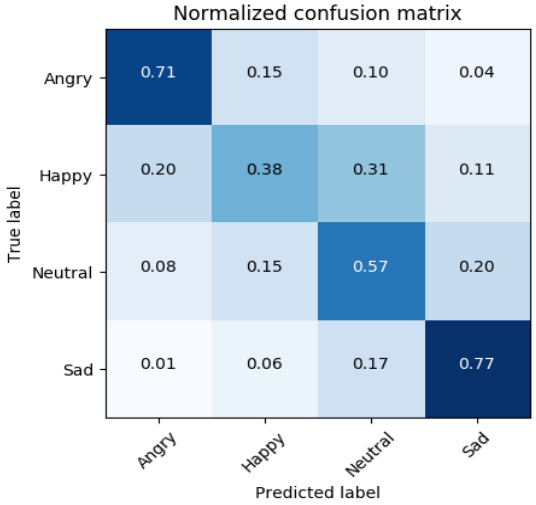
Figure 4.7 shows the scatter plot using T-sne [91] for mel spectrogram features on the training set in the SD scenario. From 4.7a, it can be seen that the distribution of 3D mel spectrograms for the four emotion classes overlaps with no clear boundaries between them. 3D AG-TFNN and Parallel AG-TFNN architecture proves successful in learning the distribution and creating clusters according to emotion classes. However, as seen from 4.7c and 4.7d, it can be concluded that the Parallel AG-TFNN architecture does a better job in separating the class-wise clusters as the overlap between the clusters can be seen to be less in Parallel AG-TFNN than in TFNN. Also, the distance between the dimension categories, i.e., valence, which comprises Sadness and Neutral, and activation, which comprises Angry and Happy, is maximized in Parallel AG-TFNN compared to TFNN.

Table 4.7: Weighted Accuracies (WA) and Unweighted Accuracies(UA) for Speaker Independent(SI) experiments using CNN+LSTM and the proposed architectures TFNN, AG-TFNN, 3D AG-TFNN, and Parallel AG-TFNN on IEMOCAP dataset.

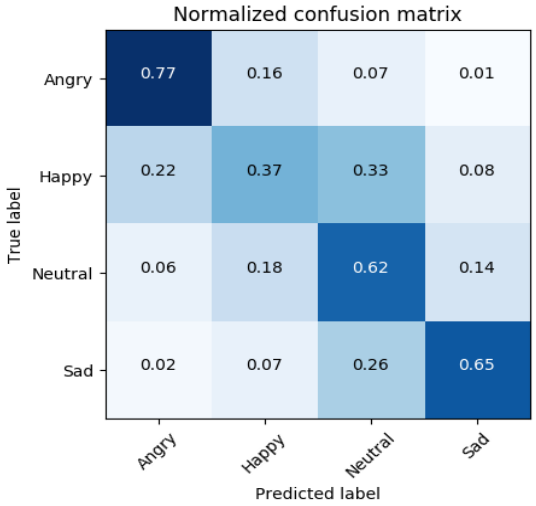
Method	Speaker Independent	
	WA	UA
CNN+LSTM	51.82(0.22%)	52.47(1.09%)
TFNN	50.21(1.61%)	51.72(1.39%)
AG-TFNN	51.42(2.00%)	52.26(1.82%)
3D AG-TFNN	53.15(2.47%)	54.07(2.84%)
Parallel AG-TFNN	**53.45(2.5%)**	**55.56(3.63%)**

4.8.2 Performance analysis using IEMOCAP

The IEMOCAP dataset comprises 4423 train utterances and 1107 test utterances belonging to the four emotions



(a) Speaker Dependent



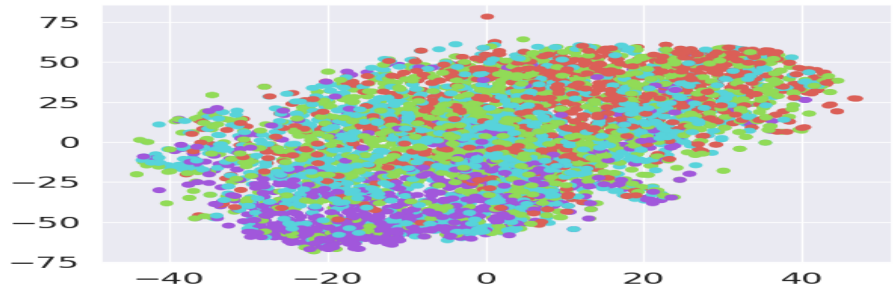
(b) Speaker Independent

Figure 4.8: Normalized Confusion matrix for Parallel AG-TFNN architecture using Mel Spectrogram and Mod Spectrogram for (a) Speaker Dependent and (b) Speaker Independent Scenario on IEMOCAP dataset

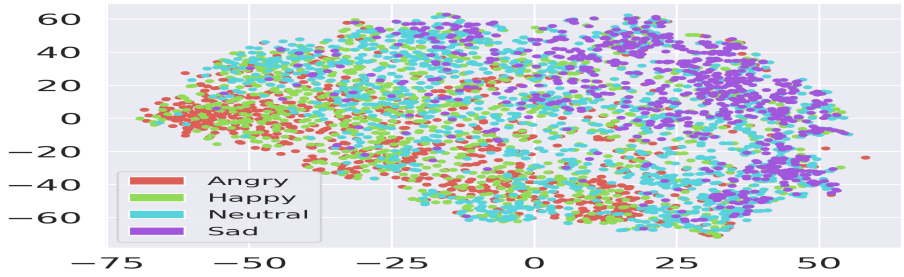
classes in the speaker-dependent(SD) case. The emotional information in IEMOCAP utterances is more natural-like than in Emo-DB; hence, more confusion can be seen in learning the model in this case. Table 4.6 and 4.7 shows the recognition accuracy of IEMOCAP dataset for the four Tensor Factorized

architectures used in the SD and SI scenario respectively. The 3D AG-TFNN and Parallel AG-TFNN architecture can generalize well using 3D mel spectrogram (for 3D case) and Mod-spectrogram features as input compared to the baseline. Figure 4.8a and Figure 4.8b shows the confusion matrices for SD and SI scenario using the Parallel AG-TFNN architecture. The Parallel AG-TFNN models the Angry and Sadness emotion categories well, but maximum confusion can be seen in modeling the Happy and Neutral class for both SD and SI scenarios.

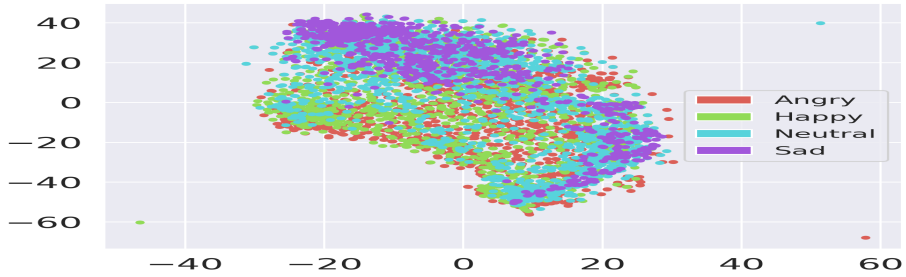
Figure 4.9 shows the T-sne scatter plot for the training data distribution of the IEMOCAP dataset. The complexity of the dataset can be visualized from 4.9a, which is the distribution of untrained data using mel spectrogram features. The four emotion classes overlap to a large extent, making the SER task on IEMOCAP more difficult than Emo-DB. Moreover, 4.9c and 4.9d show the distribution after training using 3D AG-TFNN and Parallel AG-TFNN architecture, respectively. The T-sne scatter plots show that the Parallel AG-TFNN architectures are more successful in clustering the valence and activation classes. However, the limitation arises when classifying the Happy and Angry sub-classes of the activation class and the Neutral and Sadness of the valence class. The cause can be attributed to the IEMOCAP dataset being more naturalistic than acted datasets like Emo-DB.



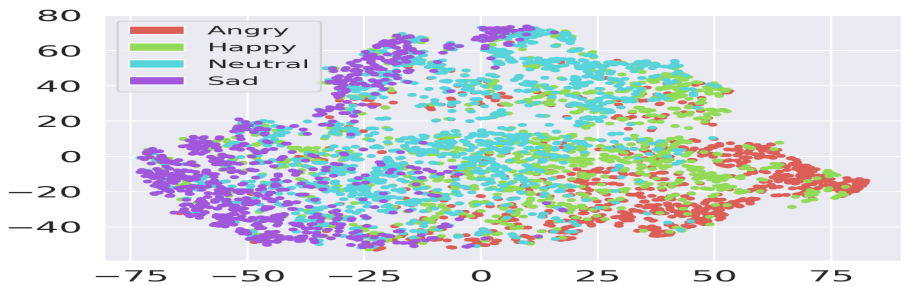
(a) Untrained distribution of emotions



(b) Embeddings from trained CNN+LSTM network



(c) Embeddings from trained 3D AG-TFNN network



(d) Embeddings from trained Parallel AG-TFNN network

Figure 4.9: T-sne scatter plot showing the distribution of the training set of IEMOCAP using Mel-spectrogram as input for the four emotion classes. 4.9a displays the untrained distribution, 4.9b displays the distribution after training by CNN+LSTM architecture, 4.9c displays the distribution after training by 3D AG-TFNN architecture and 4.9d displays the distribution after training by Parallel AG-TFNN architecture.

Table 4.8: Weighted Accuracies (WA) and Unweighted Accuracies(UA) for Speaker Independent(SI) experiments using IEMOCAP dataset for 2D and 3D AG-TFNN with delta and double delta features included in 2D and 3D form, respectively.

Method	Accuracy	
	WA	UA
2D AG-TFNN	50.95(1.80%)	52.44(2.67%)
3D AG-TFNN	53.15(2.47%)	54.07(2.84%)

4.8.3 Comparison of 2D and 3D mel spectrograms with delta features

The improvement in recognition performance of 3D AG-TFNN architecture over 2D feature-based architectures such as CNN+LSTM, 2D TFNN, and 2D AG-TFNN is due to the multi-way information capturing capability of 3D AG-TFNN in all three modes of the 3D mel spectrogram tensor rather than because of inclusion of delta and double-delta. Table 4.8 shows the comparison of recognition accuracy when delta and double delta features are stacked frame-wise along the columns of the mel-spectrogram. The resultant 2D feature contains delta and double-delta features that keep the 2D shape intact. Comparing 3D AG-TFNN with 3D log mel spectrogram features, it is evident that placing the delta and double-delta as the third mode helps capture multi-way information, thereby contributing to the improvement in recognition performance.

4.8.4 Performance Comparison With State-of-the-Art Methods

Several research works have used Emo-DB dataset for performance evaluation of the methods in SER. In [92], utterance level feature vectors were classified using a pre-trained SVM and LDA classifier, achieving an accuracy of 80% for seven class classifications on Emo-DB. Moreover, the validation accuracy reported for Emo-DB 7 class speaker-independent classification is 82.42% in [26], which uses a 2D CNN+LSTM architecture to model the emotions. Another work in [93], reported accuracies of 79.19% to 87.49% using several architectures and feature-fusion and decision-level fusion techniques. Using 3D Log-mel spectrograms and an attentive CNN+BLSTM architecture, [27] reported an accuracy of 82.82% and with 2D log-mel spectrograms, the performance reported was 79.38%. In our work, we have chosen four emotions from both Emo-DB and IEMOCAP datasets, so that comparison can be derived for the modeling capability in both the dataset scenarios, which are pretty diverse.

For the IEMOCAP dataset, [93] reported accuracies in the range 51.44% to 57.99% using architec-

Research work	Method	WA	UA
[92]	SVM+LDA	-	80 %
[26]	2D CNN+LSTM	-	82.42 %
[93]	Different architectures Feature/decision Fusion		79.19% to 87.49%
[27]	2D Mel spectrogram, 2D CNN+LSTM+Attention		79.38 %
[27]	3D Mel spectrogram, 3D CNN+LSTM+Attention		82.82 %
proposed	**3D AG-TFNN**	**85.56(6.06%)**	**85.15(6.45%)**

Table 4.9: Comparison with State-of-the-art techniques on Emo-DB dataset with speaker-independent setting

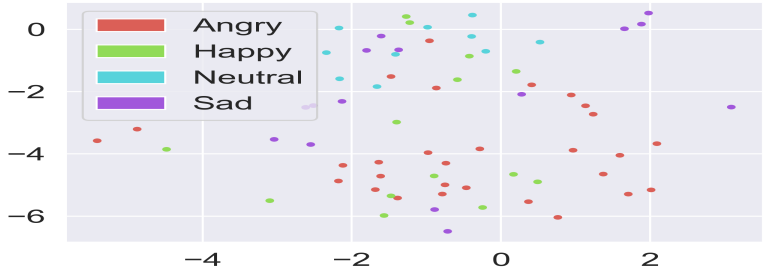
Research work	Method	WA	UA
[93]	DNN-ELM, DNN-KELM, CNN-BLSTM etc	-	51.44% to 57.99%
[94]	Data Augmentation	-	56.0 %
[26]	2D CNN+LSTM	-	52.14 %
proposed	**3D AG-TFNN**	**53.15(2.47%)**	**54.07(2.84%)**
	Parallel AG-TFNN	**53.45(2.5%)**	**55.56(3.63%)**

Table 4.10: Comparison with State-of-the-art techniques on IEMOCAP dataset with speaker-independent setting

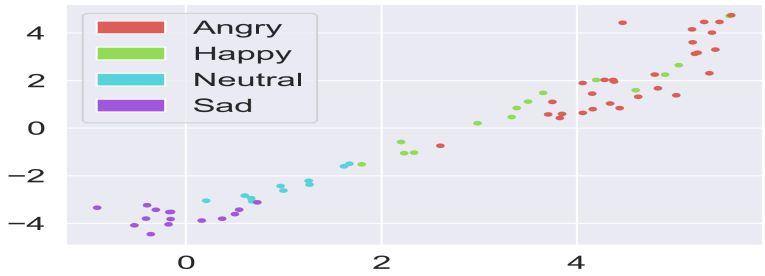
tures such as DNN-ELM, DNN-KELM, CNN-BLSTM, CNN-ELM, CNN-KELM and others including feature fusion, when the entire dataset, both scripted and improvised scenarios, were considered for the study. However, the work in [26] used only those utterances for which at least three annotators agreed for six emotion classes. They reported validation accuracy of 52.14% in the speaker-independent case. The architecture used was 2D CNN+LSTM, and log-mel spectrograms were used as input. Also, in [94], data augmentation was used to address the problem of data imbalance, and UAR of 56% was reported for the whole dataset. Compared to these methods, our proposed Tensor Factorization-based architectures outperforms many of these techniques, with the added advantage of fewer parameters and less complex architecture.

4.8.5 Comparison of GeMAPs feature and AG-TFNN embeddings for SER

It has been discussed in literature that a set of speech features such as GeMAPs [15], ComPARe [95], IS09/13 [12] feature set etc. captures paralinguistic information such as emotional content in speech. To demonstrate the effectiveness of the proposed AG-TFNN embedding based speech features against the traditional feature sets, we compare the distribution of unseen test set for both AG-



(a) Distribution of Emotions using geMAPs Feature Set



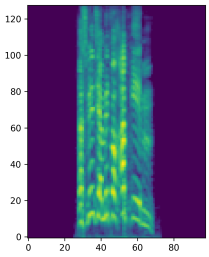
(b) Distribution of Emotions using AG-TFNN Embeddings

Figure 4.10: T-sne scatter plot showing the distribution of the test set of Emo-DB using geMAPs Feature set and AG-TFNN embeddings for the four emotion classes. 4.10a displays the distribution for geMAPs features, and 4.10b displays the distribution for embeddings calculated from pre-trained AG-TFNN.

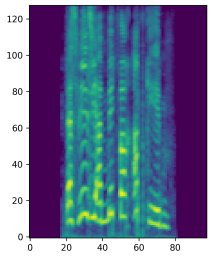
TFNN embeddings and 88-dimensional GeMAPs feature set which comprises of higher order functions calculated from Low-Level Descriptors(llds) such as pitch and its harmonics, MFCC etc. We utilized the test split of the Emo-DB dataset and calculated geMAPs functional of 88 dimensions for each speech utterance as well as embeddings from trained AG-TFNN model. Furthermore, the embeddings are flattened to a vector and T-sne is used to calculate 2-dimensional vectors, which are used to generate scatter plots as shown in Figure 4.10. It is clearly observed from the scatter plots that the embeddings generated from AG-TFNN are much more effective in capturing emotion-related information compared to the standard speech feature sets such as geMAPs.

4.9 Discussion

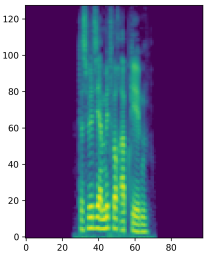
To further analyze the information being captured using AG-TFNN, a comparison of the feature tensors obtained from the hidden layer of the AG-TFNN network and feature maps obtained from the hidden layer of the baseline CNN+LSTM network is shown in Figure 4.11. For obtaining the



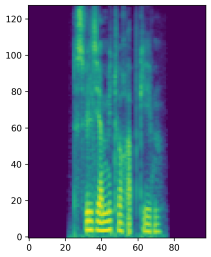
(a) Angry speech



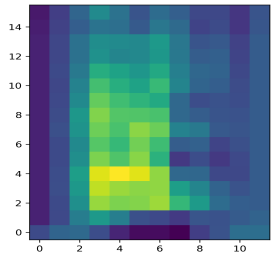
(b) Happy Speech



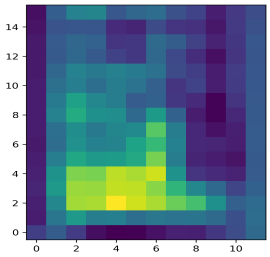
(c) Neutral Speech



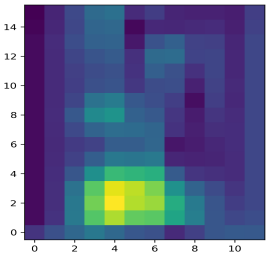
(d) Sad Speech



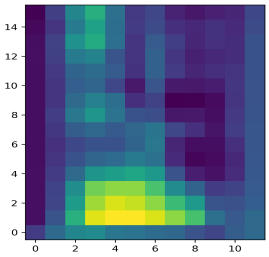
(e) CNN Feature for Angry Speech



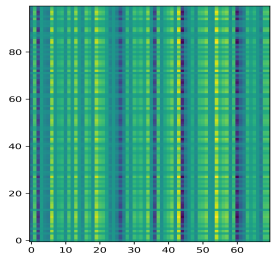
(f) CNN Feature for Happy Speech



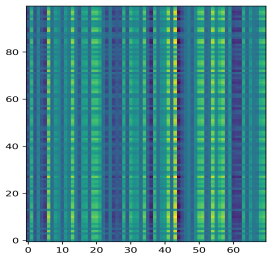
(g) CNN Feature for Neutral Speech



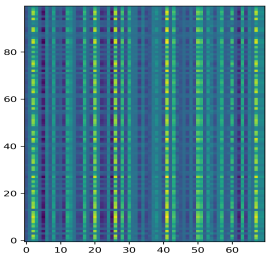
(h) CNN Feature for Sad Speech



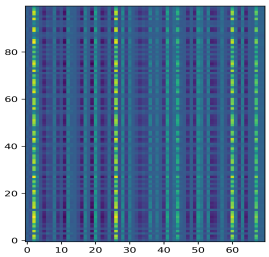
(i) Feature Tensor for Angry speech



(j) Feature Tensor for Happy Speech



(k) Feature Tensor for Neutral Speech



(l) Feature Tensor for Sad Speech

Figure 4.11: Figure (a)-(d) shows the Mel Spectrogram representation of utterances of Emo-DB dataset for the four emotion classes- Angry, Happy, Neutral, and Sad. Figure (e)-(h) represents the feature maps obtained from a randomly chosen filter from the fourth LFLB of baseline CNN+LSTM architecture for the four emotional utterances. Figure (i)-(j) represents the feature tensors obtained from the fourth Tensor FF layer for the four emotional utterances.

feature maps from the fourth LFLB of baseline CNN+LSTM, the input mel spectrogram is passed through the trained network, and feature maps corresponding to the 128 filters of the fourth LFLB are obtained for all four utterances. Now a filter is randomly selected out of the 128 filters, and a feature map corresponding to that randomly chosen filter is shown in Figure 4.11. Similarly, for obtaining the feature Tensor from the AG-TFNN network, the four mel spectrogram of utterances corresponding to the four emotion classes are passed through the trained AG-TFNN network, and the feature tensor corresponding to the last Tensor FF layer is obtained.

It can be observed that the CNN filters tried capturing the energy concentration in the mel-spectrogram, as can be seen from (e)-(h) of the Figure 4.11. The yellow region in the feature maps exhibits energy concentration learned by the CNN filter. The yellow region is more spread out towards higher frequencies in the case of Angry and Happy utterances. However, for Neutral and Sadness utterances, the yellow region can be observed to be concentrated in the low-frequency regions. In contrast, the feature tensors obtained for the four emotions using AG-TFNN exhibit differences because different neurons are activated with different intensities. Thus, different emotions involve tensor projection using different combinations of columns from the factor matrices. Also, for emotions such as Angry (Figure 4.11(i)) and Happy (Figure 4.11(j)), which involves higher energy in the higher frequency regions too, it can be observed that the feature tensors are denser, implying more number of rank-1 tensors combining to form the feature tensor. Nevertheless, in contrast, for emotions like Neutral (Figure 4.11(k)) and Sadness (Figure 4.11(l)), where the energy is mainly concentrated in lower frequency regions, the feature tensors obtained are relatively sparse, giving an implication that less number of rank-1 tensors combine to form feature tensors in such cases.

4.10 Conclusion and Future Work

In this work, new tensor factorization-based architectures are introduced for the task of SER from 2D and 3D representations of speech, such as Tensor Factorized Neural Networks (TFNN) and 2D and 3D Attention-Gated TFNN (AG-TFNN) and Parallel AG-TFNN. The 2D representations such as mel spectrograms form a natural tensor of second order, and TFNN becomes the apt choice to keep the tensor form of the input intact. Moreover, 3D Log Mel Spectrograms have found recent interest amongst the researchers for SER. To assess the modeling capacity of TFNN and AG-TFNN, experiments were conducted on two datasets - Emo-DB and IEMOCAP and the recognition rates

were compared with the baseline CNN+LSTM architecture. 3D AG-TFNN(for Emo-DB) and Parallel AG-TFNN (for IEMOCAP) for SER reached the state-of-the-art in fewer parameters and less computational complexity than the baseline. Moreover, an analysis of the learned deep features from Baseline CNN+LSTM and AG-TFNN is also presented to understand the architectures' emotion-capturing capability. AG-TFNN can be explored as an alternative to CNN+LSTM architectures for use with data with an inherent tensor structure.

5

Robust Cross-Cultural SER Leveraging Speaker and Language cues

Objective

Speech Emotion Recognition(SER) has been an active area of research in order to make Human-Computer Interaction (HCI) smoother and more natural. However, due to the dependence of the expressed emotions in an utterance on factors like culture, speaker, etc., the robustness of the SER systems in the multi-cultural setting is always a topic of discussion among researchers. Both universality and cultural specificity of emotions are debated in the literature. Thus, we propose two methods, one incorporating cultural specificity and another demonstrating the universal nature of emotions across cultures. In this work, we propose a novel method to make a multi-cultural SER by incorporating impactful factors such as speaker and language as markers of cultural distinctiveness. We develop a language and a speaker model to get language and speaker embeddings, and a multi-modal fusion architecture is proposed to fuse the information along with emotional cues. Moreover, a triplet-loss-based multi-cultural SER is proposed, which tries to normalize speaker and cultural variabilities and focuses on learning emotions, irrespective of culture. Experiments conducted on a collection of five language emotion dataset shows the robustness of the proposed technique in predicting emotions in leave-one -language-out setting. The system's design allows for the incorporation of a new language and speaker without retraining the whole system again.

5.1 Introduction

Speech is the most preferred form of communication between individuals as it contains a multitude of information such as message, which is the primary objective, and abstract information such as speaker, gender, emotional states, etc. Speech Emotion Recognition (SER) is the method of identifying and classifying emotional states underlying a speech utterance. With the evolution of Human-Computer Interaction (HCI), researchers have picked up an interest in improving the performance of Speech Emotion Recognition (SER) with the motivation to make HCI more natural and smooth and to behave according to the mood of the instructor. This becomes particularly significant with technologies such as Alexa, Cortana, etc., used in everyday life. Moreover, SER has found potential application in several important areas such as mental health [1], education [96], military [97], business [98] to name a few. However, SER still poses challenges when it comes to cross-cultural and cross-corpus settings. The recognition performance of SER systems is heavily influenced when trained with one language and tested with another. Even in the same language setting, the performance highly

degrades when trained with one corpus and tested with another. Moreover, within a single corpus, the emotional states underlying an utterance vary significantly with the speaker. As such, the performance of the speaker-independent (SI) system is generally lower than the speaker-dependent (SD) system. Earlier SER studies have primarily focused on training and testing within the same corpus or language [28]. Cross-corpus poses a challenge for researchers due to the incorporation of variabilities such as speakers, age of speakers, personality, language, labeling technique, recording environment, etc. Thus, designing an SER system robust to these variabilities is the need of the hour so that the system is adaptable to diverse languages and performs better in real-time scenarios.

The way of expressing emotions is significantly influenced by the culture. There is debate amongst researchers over the universal nature of emotions versus the cultural specificity of emotions [99]. However, few recent studies have considered both the universality and the cultural specificity of emotions [100]. Traits such as native language and cultural background are determined by region, upbringing, and culture and thus cannot be measured directly. However, a speaker's first language and its varieties are considered one of the prominent markers of culture [101]. Moreover, language cues can be used as information to aid the task of multi-cultural adaptivity of the SER models. Also, the speaker's personality contributes to the inter-speaker emotion variabilities. Training and testing the models in a speaker-independent setting is one way to get around it. However, due to the limited number of speakers in emotion datasets, it is impossible to incorporate all the intra-class variabilities for the emotion classes. Hence, the need to incorporate speaker ques into the SER model for adaption to individual speakers becomes prominent.

The area of SER has seen a dramatic shift with the introduction of deep learning techniques. The traditional method for SER consisted of extracting frame-level handcrafted features and training classifiers such as GMM, SVM, HMM, etc., intending to classify the utterances into emotional classes. With the surge in the deep learning domain due to the increase in available data and computational power, the SER systems' performance has significantly improved. Recent studies have leveraged deep architectures to extract deep features from speech representations such as mel-spectrograms [102], low-level descriptors (lld's) such as pitch and its harmonics, formant frequency, etc. extracted using standard feature sets such as IS09/IS13-feature set, ComParE, GeMaps, etc. [103]. In [27], 2D and 3D log-mel spectrograms are used in conjunction with Attentive CNN+BLSTM network for emotion classification on standard datasets of German and English language such as [104] and IEMOCAP

[105] respectively. The work showed that the models successfully captured the emotions in both languages, but no cross-corpus study was considered. Moreover, recent works like [28] and [93] have investigated several CNN and LSTM-based architectures for the task of SER. However, works such as [31], [29], [26] also investigated deep feature extraction directly from raw speech signals instead of using time-frequency speech representations.

5.2 Related Work

Most of the work in literature dealing with cross-cultural or cross-corpus studies is centered around training with one language/corpus and testing with another. Several works have studied the impact of linguistic ability as well as the influence of language on the recognition performance of emotional states underlying an utterance. In [106], a domain-adaptive version of Least-square regression is proposed to mitigate the mismatch between source and target speech datasets. Moreover, additional unlabelled data from the target speech corpus is used to augment the labeled data from the source speech corpus and is jointly trained in the transfer learning framework. Also in [107], a two-stage classification system is proposed. In the first stage, individual languages are modeled separately to preserve cultural traits. In the second stage, the Emotion Profile (EP) technique maps emotions from a never-seen language to a tractable space for classification. Moreover, transfer learning is the most sought-after strategy when it comes to cross-corpus SER. As such, [108] introduced a novel transfer linear subspace learning framework, both supervised and unsupervised, for learning a feature subspace common to both the source and target corpus. Corpus similarity was measured using the nearest neighbor graph, and a feature selection strategy was used to group the emotional features into two parts - high transferable and low transferable. Also, in [109] a feature selection strategy was proposed to select emotional acoustic features irrespective of language and other factors. The proposed strategy showed comparable results with full feature sets on various emotional corpora. Moreover, [110] explored the cross-lingual scenario when the system is trained on western languages such as English, Italian and German and tested on an unseen language Urdu. They concluded that incorporating datasets of many languages while training boosts testing accuracy for unseen Urdu datasets. Also, leaking a small fraction of data from the unseen language helps in improving the recognition performance. In another similar study in [111], cross-lingual SER is performed using English and French. Fine-tuning of a pre-trained cross-lingual model is explored using very few samples from the target language. Recent work in [112]

utilized multi-task learning to simultaneously learn emotion and language ID to incorporate language-specific information.

All these studies on cross-corpus scenarios point towards two aspects - universality of emotions across cultures and cultural specificity of emotions. This motivated us to explore both the aspect of cross-cultural SER - one by additionally incorporating language and speaker variability and the other by normalizing these variabilities. A detailed description and analysis of the methods are mentioned in the subsequent chapters. However, a summary of the significant contributions of this chapter are as follows -

(i) We propose a novel multi-modal SER system using audio modalities such as emotion, language, and speaker for a multi-cultural scenario.

(ii) Building upon the motivation that emotions are universal across cultures, a triplet-loss-based metric learning approach is also investigated to normalize cross-cultural and personality of speakers-related variabilities.

(iii) Embeddings extracted from a metric-learning-based model can be combined with a simple DNN classifier to predict labels for unseen languages scenario.

(iv) Experimental evaluation of emotion datasets belonging to five different languages shows that the proposed techniques are robust enough to tackle the challenge of cross-cultural variabilities.

5.3 Dataset

The data requirement to train a deep learning model is relatively high, and getting an SER dataset containing multiple languages and speakers is difficult. For this study, we used five SER datasets belonging to five languages to incorporate cultural and speaker variability. The datasets are recorded in different conditions and, as such, pose a challenge for cross-corpus analysis. A brief description of the five datasets is provided below.

5.3.0.1 EmoDB- German Emotional Dataset

The Berlin Emotional Database(Emo-DB) [104] is a German emotional speech corpus consisting of seven emotion categories, i.e., Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral, recorded from ten German actors. The speech utterances were recorded using a sampling rate of

Table 5.1: Distribution of utterances among the emotion classes for the Five emotion datasets

Dataset	Language	Num utterances/ duration	Num speakers
Emo-DB [104]	German	535 utterances	10
eNTERFACE [113]	English	1166 utterances	42
IITKGP-SEHSC [114]	Hindi	7000 utterances	10
IITKGP-SESC [115]	Telegu	12000 utterances	10
Shemo-DB [116]	Persian	3000 utterances	87

16 kHz with a 16-bit resolution and mono-channel, and the entire dataset comprises 535 sentences. The average duration of the audio files is three seconds. We selected the four basic emotions- Anger, Happiness, Neutral, and Sadness for our experiment purpose.

5.3.0.2 Enterface - English Emotional Dataset

The eNTERFACE'05 Audio-Visual Emotion Database [37] consists of 1166 video sequences presented by 42 subjects (coming from 14 different nationalities, 81% men and 19% women). Each subject was asked to listen to six different stories eliciting particular emotional states: Anger, Disgust, Fear, Happiness, Sadness, and Surprise. After that, the subject read five utterances in English, constituting five different reactions to the given situation. Two human experts assessed all samples.

5.3.0.3 IITKGP-SEHSC - Hindi Emotional Dataset

The database is recorded using professional artists from Gyanavani FM radio station, Varanasi, India. The speech corpus is collected by simulating eight different emotions using neutral (emotion-free) text prompts. The emotions present in the database are Anger, Disgust, Fear, Happiness, Neutral, Sadness, Sarcastic, and Surprise. This speech corpus is named as Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC). The quality of the emotions expressed in the database is evaluated using subjective listening tests. The emotion recognition performance using subjective listening tests is observed to be around 74%. The results of subjective listening tests are grossly on par with the results obtained using a prosodic analysis of the database.

5.3.0.4 IITKGP-SESC - Telugu Emotional Dataset

The database is recorded using ten (five male and five female) professional artists from All India Radio (AIR) Vijayawada, India. All the artists are 25-40 years old and have professional experience

Table 5.2: Genetic proximity of languages considered for this study calculated using eLinguistics.net language comparison tool. A value of 0 means the two languages under comparison are the same, whereas a value of 100 means that the two languages under comparison are unrelated.

Languages	German	English	Hindi	Telugu	Persian
German	0				
English	30.85	0			
Hindi	76.5	65.2	0		
Telugu	83.4	93.1	93.3	0	
Persian	86.7	78.0	55.9	92.7	0

of 8-12 years. The speech corpus is collected by simulating eight different emotions using Neutral (emotion-free) statements. Each artist has to speak the 15 sentences in eight basic emotions in one session. The number of sessions considered for preparing the database is ten. The total number of utterances in the database is 12000 (15 sentences × eight emotions × ten artists × ten sessions). Each emotion has 1500 utterances. The number of words and syllables in the sentences varies from 3-6 and 11-18, respectively. The total duration of the database is around 7 hours. The eight basic emotions considered for collecting the proposed speech corpus are Anger, Compassion, Disgust, Fear, Happy, Neutral, Sarcastic, and Surprise.

5.3.0.5 Shemo-DB - Persian Emotional Dataset

ShEMO-DB is a large-scale, validated database for Persian called Sharif Emotional Speech Database. The database includes 3000 semi-natural utterances, equivalent to 3 h and 25 min of speech data extracted from online radio plays. The ShEMO covers speech samples of 87 native-Persian speakers for five basic emotions, including Anger, Fear, Happiness, Sadness, and Surprise, as well as a Neutral state. Twelve annotators label the underlying emotional state of utterances, and majority voting is used to decide on the final labels. According to the kappa measure, the inter-annotator agreement is 64%, which is interpreted as "substantial agreement". We also present benchmark results based on common classification methods in speech emotion detection tasks. According to the experiments, support vector machine achieves the best results for both gender-independent (58.2%) and gender-dependent models (female = 59.4%, male = 57.6%).

Table 5.3: Mean Recognition Accuracies in terms of Weighted Accuracy (WA) and Unweighted Accuracy (UA) for the individual datasets in a five-fold cross-validation setting .

Method	German		English		Hindi		Telegu		Persian	
	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
CNN+LSTM	91.76	90.95	89.41	89.67	87.99	87.84	83.07	83.22	73.54	57.71
2D AG-TFNN	86.47	84.80	81.76	82.50	70.85	70.95	72.31	72.40	68.24	49.96
3D AG-TFNN	92.40	90.55	88.47	88.20	86.23	87.00	84.12	83.56	70.89	55.22

5.3.1 Genetic proximity comparison

Since this study considers both cultural specificity and universality, we calculated the genetic proximity of the languages to asses how emotions are related when two languages are genetically proximal. To evaluate the genetic proximity of the languages, we utilized the online tool eLinguistics.net [117]. For two languages under comparison, it generates a score between 0 and 100, with 0 being the same language and 100 being the two languages unrelated. The methodology consists of choosing language material for the comparisons (basic words in this case). A total of 18 words have been chosen, often used in comparative linguistics studies. The next step is cognate scoring, which compares the consonants contained in a word in a language and the corresponding word in the language to which it is being compared. The order in which these consonants appear in words is being considered. Finally, statistical context is calculated out of the cognate scores.

5.4 Base Architectures

The proposed methodologies make use of the base architectures for modeling emotions. We investigated the use of CNN+LSTM [26] architecture, whose variants are widely used in SER studies and 2D and 3D Attention Gated Tensor Factorized Neural Networks (AG-TFNN) [118]. The motivation for using AG-TFNN architectures is that mel-spectrograms, which are inputs to the models in our work, are naturally in 2D tensor form. Also, recent works such as [118] have shown that 3D log-mel spectrograms outperforms their 2D counterpart and are in third-order tensor form. Hence, to reduce the network's complexity and the number of parameters and training time, we chose to use AG-TFNN alongside the popular CNN+LSTM framework. A brief description of both CNN+LSTM

and AG-TFNN architecture is mentioned below.

5.4.1 CNN+LSTM Architecture

The CNN+LSTM architecture used in our work is adapted from [26]. The architecture is divided into two parts - Local Feature Learning Blocks (LFLBs) to extract local patterns from speech spectrograms and Global Feature Learning to aggregate utterance level features by capturing the temporal dependencies, which is of significance in SER studies.

Each LFLB comprises one convolutional layer, one batch-normalization layer, one activation layer, and one max pooling layer. CNN performs local feature learning using 2D spatial kernels. CNN has the advantage of local spatial connectivity and shared weights, which helps the convolution layer perform kernel learning. Batch Normalization is performed after the convolution layer to normalize the activations of each batch by maintaining the mean activation close to zero and standard deviation close to one. The activation function used is Exponential Linear Unit (ELU). Contrary to other activation functions, ELU has negative values too, which pushes the mean of the activations closer to zero, thus helping to speed up the learning process and improving performance [83]. Max pooling is used to make the feature maps robust to noise and distortion and also helps reduce the number of trainable parameters in the subsequent layers by reducing the size of the feature maps.

Global feature learning is performed using an LSTM layer. The output from the last LFLB is passed on to an LSTM layer to learn the long-term contextual dependencies. Sequences of high-level representation obtained from the CNN+LSTM architecture are passed on to an attention layer whose job is to focus on the emotion salient parts of the feature maps since not all frames contribute equally to the representation of the speech emotion. The attention layer generates an utterance level representation, obtained by the weighted summation of the high-level sequence obtained from CNN+LSTM architecture with attention weights obtained in a trainable fashion [27]. The utterance level attentive representations are passed to a fully-connected layer and then to a softmax layer to map the representations to the different emotion classes.

The number of convolutional kernels in the first and second LFLB is 64, and for the third and fourth, LFLB is 128. The size of the convolutional kernels is 3×3 with a stride of 1×1 for all the LFLBs. The size of the kernel for the max-pooling layer is 2×2 for the first two LFLBs and 4×4 for the latter two LFLBs. The size of the LSTM cells is 128.

5.4.2 AG-TFNN Architecture

Tensor Factorized Neural Network was first introduced in [80]. It is based on Tucker Decomposition [44] of n-Dimensional tensors, preserving the multi-way relationship among the tensor modes. Further, an Attention Gated TFNN (AG-TFNN) was proposed for SER in [118], which introduced a Tensor attention mechanism to provide attentive tensor inputs to the TFNN network, keeping the tensor form of the inputs intact (both 2D and 3D tensors).

The AG-TFNN architecture consists of one Attention Layer, three Tensor Feed Forward (FF) layers, and one Tensor Softmax Layer. The Tensor Feed Forward (FF) block is based on Tucker Decomposition of Tensors, where an input tensor of order N is decomposed into a core tensor whose size is less than or equal to the original tensor and N factor matrices, one corresponding to each mode of the tensor. The core tensor provides coefficients of interaction amongst the columns of the factor matrices and thus contains information about the multi-way relationship amongst the tensor modes. The core tensor is then passed through an activation function to obtain the feature tensor pertaining to that Tensor FF layer.

The feature tensor from the last Tensor FF layer is then projected to an $N+1$ dimensional weight tensor, where the length of the $N+1$-th dimension is equal to the number of the output emotion classes, to yield class probabilities using a softmax activation. Moreover, we have used a Tensor Attention layer at the beginning of the TFNN network to provide emotion-specific attentive inputs to the TFNN network. The attentive input is obtained by projecting the input tensor on the factor matrices in the attention layer. The attention layer output is the element-wise dot product of the emotion weights with the input elements in the spectrogram, thus highlighting the emotion salient regions.

5.5 Methodology

This section discusses the different techniques employed in the task of multi-lingual SER. To this aid, we first discuss the base architectures - CNN+ LSTM and Attention Gated Tensor Factorized Neural Network (AG-TFNN)- used for our work. Next, the recognition performance of the architectures used on single-language SER datasets and the techniques used in multi-lingual SER are discussed. To incorporate the universality and cultural specificity of emotions, we utilize two techniques - the multi-modal approach using audio modalities and the generalized approach using the triplet loss model.

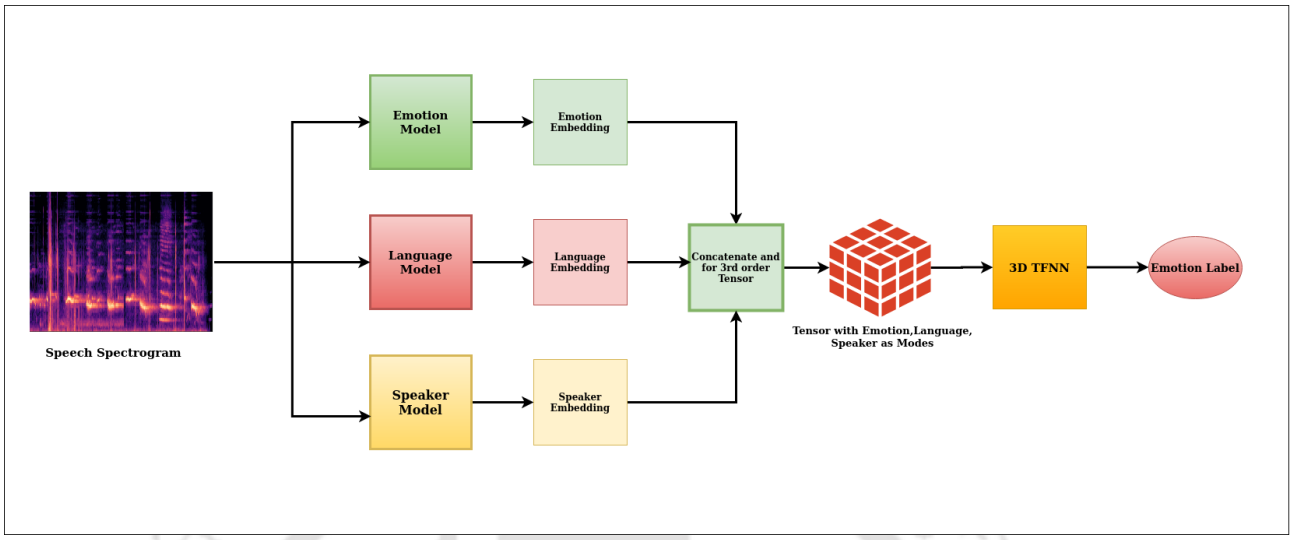


Figure 5.1: Proposed methodology using classification approach. The Individual Emotion, Language, and Speaker Model are made untrainable to extract embeddings only.

Both the techniques are discussed in detail below.

5.5.1 SER on Individual Datasets

We compute the within-language recognition accuracy by five-fold cross-validation for each language using five random seeds for the train-test set partition. The architectures used are the baseline CNN+LSTM architecture and the Attention-Gated Tensor Factorized Neural Network (AG-TFNN) with mel-spectrogram inputs for CNN+LSTM and 2D AG-TFNN and 3D mel spectrogram for 3D AG-TFNN architecture. This preliminary classification allows setting the first baselines, and these results would represent the maximum achievable scores if the system is trained and tested using the same language. Our objective is to predict emotions in a cross-cultural setting, so these results will only serve as a guideline. Table II presents the average recognition accuracies with standard deviations across five-folds for the five individual emotion datasets on CNN+LSTM and 2D & 3D AG-TFNN architectures. The performance of these architectures suggests that the chosen architectures serve the purpose of SER well.

5.5.2 Method-I: The classification model-based approach

The first method is based on the idea that emotions are hugely impacted by functional traits such as culture and personality [99]. There is an argument amongst the researchers about the universality versus cultural specificity of emotions. As such, the proposed technique also aims to incorporate such

variabilities to make the model adaptive to cultural and personality traits. Thus, the issue of the cultural distinctiveness of emotions is apprehended by imbuing language-specific cues in addition to the SER model, as language is one of the prominent markers of culture [101]. Moreover, personality information is imbued using speaker cues by virtue of deep speaker embeddings.

Figure 5.1 depicts the structure of the proposed end-to-end system for multi-cultural emotion recognition from speech using the classification architectures. Firstly, a classification model is trained using the base architectures described in section - 5.4 for Emotion classification and Language Classification tasks. Since the number of speakers is less and to remove any bias towards the speakers present in the dataset, we utilize a pretrained speaker recognition model to extract speaker-specific embeddings for imbuing speaker traits in the SER model. Below is a brief description of the three models- emotion, language, and speaker.

5.5.2.1 Language Model

The language model is trained using the five language SER datasets in a speaker-independent fashion. The recognition accuracies for the CNN+LSTM network is 100% and that using AG-TFNN is 99%. The embeddings for language are extracted from the pre-final layer of the network in the case of AG-TFNN and the Convolutional Layer of the fourth LFLB in the case of CNN+LSTM architecture. However, this language model is limited because it cannot generate discriminative embeddings for a new language. One possible solution is to train a language model with many languages, including the unseen one used for the test scenario. Figure 5.7 shows the distribution of the languages before and after the language model training. The trained distributions correspond to the embeddings extracted from the pre-final layer of the base architectures. All three base architectures can classify the languages; hence, the language embeddings generated from such models are highly discriminative.

5.5.2.2 Speaker model

Since our main objective is not speaker recognition, we employ transfer learning to extract discriminative speaker embeddings from pre-trained networks. For the speaker embeddings, we utilize a pre-trained deep speaker model trained in Mandarin and English, consisting of many speakers both in text-dependent and text-independent scenarios [119]. The network is first pre-trained using a softmax activation in the final layer with a cross-entropy loss function. Then the pre-trained model is fine-tuned with triplet loss and cosine-similarity as the metric. This technique has shown improved

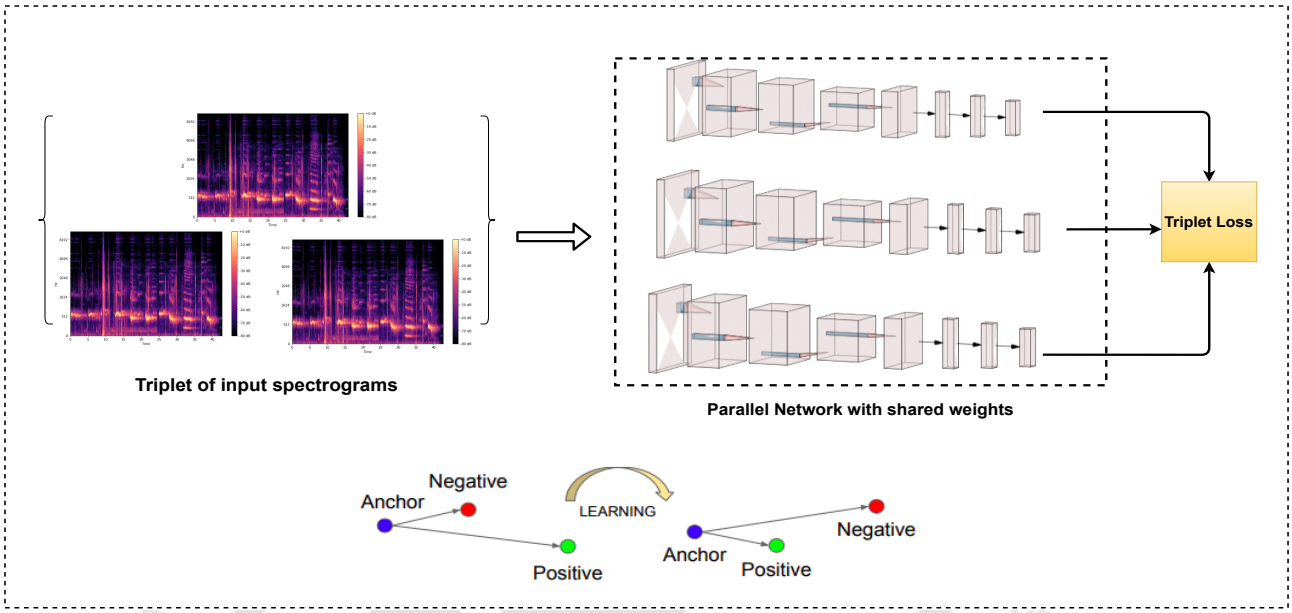


Figure 5.3: Proposed methodology using the Metric Learning approach. Triplet loss is used to compare embeddings from parallel networks with shared weights.

recognition performances. For our work, the embeddings are extracted from the pre-final layer to be fused into the multi-modal framework. The dimensions of the embedding vectors produced are 512×1. The comparison of the distribution of ten speakers from the Emo-DB dataset is shown in Figure 5.6. It shows that the pre-trained deep speaker embeddings are discriminative to a large extent and can be utilized for extracting speaker-specific cues from unseen language datasets.

5.5.2.3 Combined SER model

The input to the combined SER model is the embeddings from emotion, language, and speaker models. For models utilizing AG-TFNN as a base architecture, the embeddings in 2D/3D form are first vectorized and then mapped to a vector of dimension 512×1 using a linear layer. For models trained in CNN architecture, the embeddings are flattened and mapped to vectors of 512×1. The three embedding vectors of dimensions 512×1 corresponding to emotion, language, and speaker are then concatenated to form a feature vector and passed on to a simple DNN architecture to perform classification.

5.5.3 Method-II: The Metric Learning based approach

The second technique is based on the concept of the universality of emotions. As discussed earlier, many researchers believe that traits such as emotions in speech are universal across cultures and have

common markers in the inter-cultural scenario. Motivated by this argument, this technique aims to normalize the speaker and cultural differences rather than adapting to individual personality and culture as was in the first technique.

Earlier deep learning works in SER investigated the use of cross-entropy loss in conjunction with softmax as the supervision component, which does not explicitly encourage discriminative feature learning. It is not well suited to adapt to new unseen classes unavailable to them during training. Recent work in SER [120], and Speaker Verification [121] has investigated the use of metric learning using contrastive loss and triplet loss. These losses encourage intra-class compactness and inter-class separability between learnable features.

5.5.4 Architecture

Metric learning is an approach based directly on a distance metric that aims to establish similarity or dissimilarity between objects. While metric learning aims to reduce the distance between similar objects, it also aims to increase the distance between dissimilar objects. For this reason, there are approaches, such as k-nearest neighbors, which calculate distance information and approaches where the data is transformed into a new representation. Unlike other loss functions, such as Cross-Entropy Loss or Mean Square Error Loss, whose objective is to learn to predict directly a label, a value, or a set of values given an input, the objective of Ranking Losses is to predict relative distances between inputs.

In our work, we have employed Triplet loss which has proven to generate discriminative embeddings in classification tasks. For training the network to learn emotionally discriminative embeddings, a batch of triplets is selected using a sampling strategy. An individual triplet in the batch consists of an *anchor* x^a, which is the reference utterance, a *positive* x^p which belongs to the same class as *anchor* utterance, and a *negative* x^n which belongs to any of the rest of the classes. Neural network architecture is employed to extract d-dimensional embeddings from each utterance in the triplet sample. In our work, the base architectures - CNN and AG-TFNN are employed to extract these embeddings. The goal of this method is to bring the d-dimensional embeddings for *anchor* and *positive* sample in the triplet closer and *anchor* and *negative* sample farther in the embedding space. This is achieved by using a triplet loss function on the d-dimensional triplet embeddings, and the network parameters are updated for each batch of triplets.

Table 5.4: Distribution of Languages in train and test set for the five-fold setting.

Fold	Train Languages	Num utterance	Test Language	Num Utterance
1	German, English, Hindi, Telegu	11,934	Persian	2737
2	English, Hindi, Telegu, Persian	14,332	German	339
3	German, Hindi, Telegu,Persian	13,103	English	1568
4	German, English,Telegu,Persian	11,108	Hindi	3563
5	German, English,Hindi, Persian	8207	Telegu	6464

5.5.4.1 Triplet Loss

Since the introduction of Triplet Loss in [122] for extracting discriminative face embeddings, it has been recently explored in the SER domain in works such as [123], [124] and has shown considerable performance improvement. Given a batch of triplet of utterances containing an *anchor* x^a, a *positive* x^p and a *negative* x^n samples and f_θ be a neural network mapping the feature representations of the utterances x into a d-dimensional embeddings such that $f_\theta(x) \in \mathcal{R}^d$, the relation which the embeddings must satisfy is -

$$\| f(x_i^a) - f(x_i^p) \|_2^2 < \| f(x_i^a) - f(x_i^n) \|_2^2$$
$$\forall (x_i^a, x_i^p, x_i^n) \in \mathcal{T}$$

Where a batch of Triplets is represented by $\mathcal{T}$ and a single triplet of feature representations of utterances in the batch is represented by (x_i^a, x_i^p, x_i^n) and the similarity metric used is the *Euclidean Distance*. To maintain a sufficient distance between the positive and negative embeddings, a margin α is defined, which is empirically chosen. Therefore, the triplet loss can be formulated as.

$$L = max[0, \sum_{i=1}^{N} (\| f(x_i^a) - f(x_i^p) \|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \alpha)]$$

Where the triplet loss $\mathcal{L}$ is calculated over the mini-batch of N samples. As such, the triplet samples which are already well separated do not contribute to the loss function and hence do not affect network parameters update.

5.5.4.2 Triplet selection strategy

The triplet selection strategy highly influences the performance of triplet-loss-based networks. We have used a semi-hard-based triplet selection approach. From 5.5.4.1 , it can be seen that the loss becomes 0 for easy triplets where $\| f(x_i^a) - f(x_i^p) \|_2^2 + margin < \| f(x_i^a) - f(x_i^n) \|_2^2$ and as such wont contribute anything to parameter updates, which may result in loss of information from such samples. Moreover, if we choose only hard triplets where $\| f(x_i^a) - f(x_i^n) \|_2^2 < \| f(x_i^a) - f(x_i^p) \|_2^2$, it may affect network generalization by assigning more importance to mislabelled samples, if any. We have used a semi-hard triplet selection strategy as proposed in [122]. For a triplet batch containing N number of triplets, half of the triplets, i.e., $N/2$ is selected using hard triplet mining, and the remaining $N/2$ is chosen randomly from a large pool of pre-constructed triplets on the batch of utterances. This helps our network to learn from both the hard triplets and randomly chosen triplets, which might contain easy as well as semi-hard triplets where the distance between the anchor and negative is still larger than the anchor and positive and still have a positive loss value.

5.6 Experimental Setting

Mel-Spectrograms are used as features for both Method-I and Method-II. In Method-I, the language and emotion model is trained using mel-spectrograms. However, for extracting speaker embeddings from the pre-trained deep speaker model, MFCC is used. The utterances in different datasets are recorded using different sampling frequencies. Hence, the utterances are first downsampled to 16000 Hz. Since the input to both CNN+LSTM and AG-TFNN must be of equal length, the speech utterances are either zero-padded or chopped for 5 secs duration.

For the generation of mel-spectrograms, the equal-sized speech utterances are first windowed using a hamming window. The Short-Term Fourier Transform (STFT) is computed using a frame size of 2048 and a frameshift of 512 samples. The STFT spectrogram is mapped to mel-spectrogram using a mel-filterbank with 128 filters. The filterbank energies are passed through a log operator to get the log-mel spectrogram, which is the feature representation for our methodologies. Librosa python library [86] is used to extract the mel-spectrograms.

Methods I and II are implemented using the Keras library, with TensorFlow in the backend.

Figure 5.4: Raw MFCC

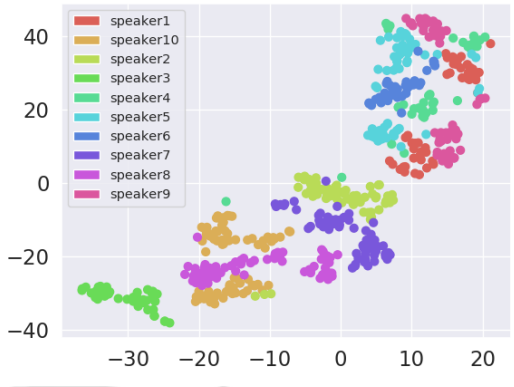


Figure 5.5: Embeddings from pre-trained network

Figure 5.6: T-sne plots showing the distribution of raw MFCC and data embeddings for pre-trained speaker model.

5.7 Results

Experiments are performed for Method I and Method II using Leave-One Language Out setting. Training is performed using four languages, and one-unseen language is used for testing. The details of the training language and test language folds are presented in Table 5.4.

5.7.1 Performance Analysis for Method-I

For Method I, a language model was first trained on the speech utterances from the five languages Emotion dataset described in Section III. The training and test partitions are speaker-independent to remove any speaker bias. For CNN+LSTM architecture, the recognition accuracy is 100%, and for AG-TFNN-based architecture, the recognition accuracy is 99.1%. Figure 5.7 shows the distribution of train and test utterances for the language model using both CNN+LSTM architecture and AG-TFNN architecture. It is evident from the figure that the language embeddings generated from these architectures are highly discriminative and class localized.

Figure 5.6 shows the distribution of untrained speakers from the Emo-DB dataset and the distribution of embeddings extracted from the pre-trained Deep Speaker Embeddings network. Even though the data distribution on which the pre-trained network was trained is very different from the data distribution of Emotion Datasets, it is evident from the figure that the embeddings produced are appreciably discriminative, thus making the model a good choice for extracting speaker embeddings.

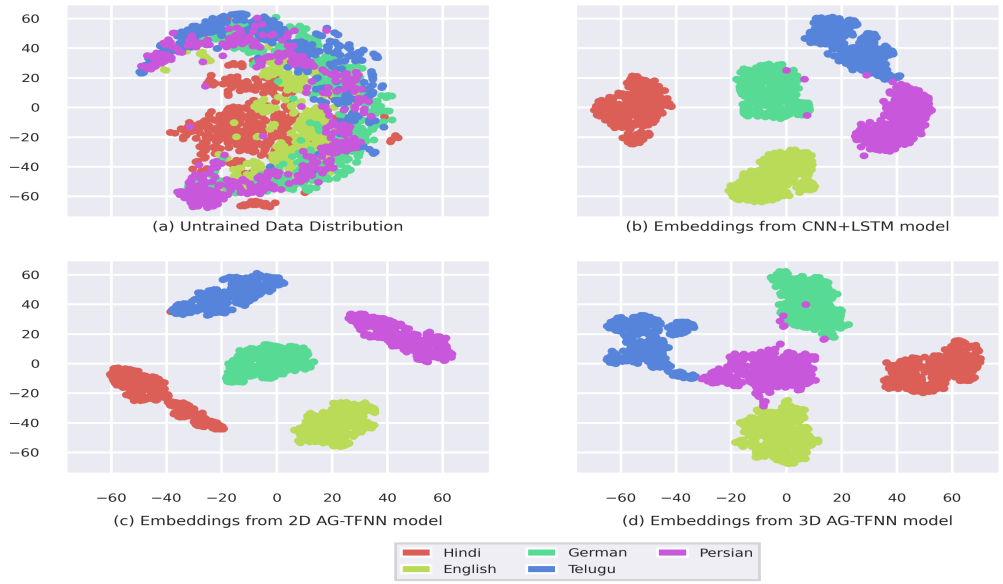


Figure 5.7: T-sne scatter plot representing the distribution of training set utterances for Language models. Figure (a) represents the untrained raw distribution, (b) represents the distribution of CNN embeddings, (c) represents the distribution from the 2D AG-TFNN model, and (d) represents the distribution of the 3D AG-TFNN model.

Table 5.5 shows the recognition performance for Method-I for all the five folds of Data as described in table5.4. The recognition accuracies for German and English language test scenario shows improvement when only Speaker embeddings are fused with emotion embeddings. The addition of Language embeddings does not affect the performance much. However, the test scenarios for Hindi, Telugu, and Persian show significant improvement with the addition of both speaker and language embeddings. Moreover, adding only speaker embeddings with emotion lowers the performance compared to the emotion-only model. This shows that certain culture-specific emotion is more affected by language markers than speaker markers.

5.7.2 Performance analysis for Method-II

For Method-II, triplet loss-based architecture is utilized for training the model in four languages. For the test case, embeddings are extracted for the speech utterances of an unseen language dataset. The embeddings serve as feature input to a single-hidden layer DNN which gives class labels of utterances. The embeddings produced from such a network are independent of cultural affects as triplet loss training tries to normalize the inter-cultural and inter-speaker variances across different language emotional corpus.

Table 5.5: Recognition Accuracies in terms of Weighted Accuracy (WA) and Unweighted Accuracy (UA) for the unseen language datasets for Method-I using CNN+LSTM as base architecture.

Test Language	Emotion Only		Emotion + Speaker		Emotion + Language + Speaker	
	WA	UA	WA	UA	WA	UA
German	91.76	90.05	92.75	**91.81**	91.17	**90.18**
English	89.41	89.67	97.93	**92.37**	95.54	**92.24**
Hindi	87.99	87.84	67.73	67.35	88.77	**88.53**
Telugu	83.07	83.22	73.76	73.67	86.85	**86.88**
Persian	73.54	57.71	75.72	59.39	76.45	**63.09**

Table 5.6 shows the recognition performance for the five folds using Method II when three different base networks are considered - 2D TFNN, 3D TFNN, and CNN network. The table shows that the 3D TFNN network as a base network is comparable with the CNN base network in a triplet loss scenario, with the added advantage of fewer parameters and computation time. Moreover, it can be observed that the triplet loss-based architecture has appreciably normalized the cultural and speaker variabilities, which can be seen from recognition accuracies on the unseen test set.

5.8 Discussion and Future Work

As can be observed from Section VII, both Methods-I and II are successful to some extent in mitigating the issues arising out of cross-cultural scenarios. However, a significant drawback with Method-I is the generation of language embeddings for a new unseen language. The language model used for Method-I has been explicitly trained for languages that were to be used for the cross-cultural emotion recognition study. Incorporating a new unseen language will require fine-tuning the language model for that new language to generate language embeddings. This issue is somewhat mitigated in the case of Method-II, where the triplet loss-based network tries to normalize the speaker and cultural variabilities, thus producing embeddings independent of language and speaker variances. As such, Method-II becomes the apt choice when it is not feasible to separately fine-tune the language model

Table 5.6: Recognition Accuracies in terms of Weighted Accuracy (WA) and Unweighted Accuracy (UA) for the unseen language datasets for Method-II using 2D TFNN, 3D-TFNN, and CNN as base architecture.

Folds	2D-TFNN base		3D TFNN base		CNN base	
	WA	UA	WA	UA	WA	UA
Fold1	72.08	55.52	75.00	57.27	75.36	**57.78**
Fold2	85.29	79.92	88.23	**87.73**	91.17	86.36
Fold3	93.94	89.32	91.71	85.32	94.26	**90.07**
Fold4	71.10	70.51	78.82	78.79	81.48	**81.02**
Fold5	73.24	73.09	83.21	**83.34**	80.43	80.19

due to the low availability of resources, as in the case of low-resource languages.

The present study is confined to only five languages, which limits its generalizing capacity to other languages. A more detailed study for cross-corpus and cross-lingual generalization can be performed as future work to extend emotion recognition for low-resource languages. Moreover, due to the application of emotion recognition from the speech in areas such as mental health, defense, etc., it becomes imperative for the models trained to be universal across cultures due to the non-availability of mental health-related data for all the languages. As such, the scope of this study can be extended to automatic diagnosis and assessment of mental health issues from speech.

6

Depression Detection from Speech using Tensor Based Multiple Instance Learning

Objective *Depression is one of the significant mental health issues affecting all age groups globally. However, its diagnosis is primarily dependent upon trained psychologists using conventional means such as interview assessment and PHQ-8/9 scores rendered manually based on the experience of the psychologists. In an era of smart wearable devices and smartphones, passive monitoring of depression traits using behavioral signals such as speech is the key to moving forward. Therefore, automatic depression classification becomes the need of the hour with quality advancements in deep learning. We propose a speech-based depression recognition system to aid and speed up the diagnosis by psychologists. Conventional audio-based techniques have utilized the Multiple Instance Learning (MIL) framework for depression label generation for individual speakers based on a set of utterances by the speaker. We propose a novel tensor-based approach in place of MIL, which requires a more straightforward and less complex architecture and extracts discriminative features for depression recognition. Experiments performed on the DAIC-WOZ depression dataset for audio modality yield state-of-the-art recognition performance.*

6.1 Introduction

Depression is a mental health issue often characterized by low mood, sadness, negative thoughts, mental disturbance, loss of interest in day-to-day activities, and is often caused by an individual's inability to cope with stressful events [125]. According to a report by World Health Organization (WHO), clinical depression is one of the primary causes of disability [126]. Moreover, severe depression cases place an individual at a higher risk of suicidal behaviour [127] as several studies have shown that people who often commit suicide meet the criteria for clinical diagnoses of depressive illness. As such, diagnoses and treatment of clinical depression become of paramount importance. Since there is no clinical characterization of depressed individuals available, the diagnoses of clinical depression often involve clinical interviews by psychologists and using the standard Hamilton Rating Scale, PHQ. -8/9 rating system to give a depression score per individual [128], [129]. However, this method is very subjective and time-consuming as well. Moreover, the reliance on a psychologist's ability to mark someone as depressed or not creates a bias.

To overcome subjective bias, automatic diagnoses and screening of depressed individuals from various biomarkers have been practiced in the medical field. Biomarkers such as low levels of serotonin [130], low functioning of neurotransmitter gamma-amino butyric acid (GABA), etc., which has

shown strong correlates for mental health-related issues [131]. However, these markers are not popular because of their invasive approach. Depression detection in individuals has to be more active and non-invasive, and thus researchers identified several behavioral markers to recognize individuals with depression. The ones widely studied are speech signal-based depression recognition [125], eye movements [132], facial activity [133], gesturing [134], slumped posture [135], etc. These markers help in automatic diagnoses of depression without disturbing the patient. They can be employed in wearable smart devices such as smartwatches, smartphones, etc., to monitor the individual's mental state continuously.

Depression recognition from behavioral signals such as speech, facial expressions, etc., has attracted researchers due to its challenging nature. Several works have investigated features and learning strategies for depression recognition from speech. The work in [136] investigated the effect of segment level and prosodic features on the classification of depressed speech from normal controls. It was pointed out that statistical functionals computed from low-level features lose information resulting in inferior performance to segment-level features. However, the work in [137] explored speech style as an aspect of depressed vs. normal speech with gender classification as a precursor to improving the recognition performance. It was found that several speech features such as MFCC, intensity, and energy features were of significance when both male and female participants' speech was considered. However, shimmer and RMS energy features were prominent for female-only depression classification, and voice quality was the marker for male participants. An investigation of temporal features revealed that the response time and average syllable duration were longer in depressed subjects. In contrast, healthy controls' interaction involvement and articulation rate were higher. Also, in [138], several speech types such as read speech, interviews, and picture description and emotion types such as positive, negative, and neutral were investigated for their discriminative power when it comes to depression versus normal speech classification. Experiments on a dataset of 74 subjects using an SVM classifier revealed that interview speech and neutral emotion contribute more to recognizing depression from speech than other speech and emotion types. Furthermore, the research in [139] investigated the effect of speaker normalization for depression classification performance as mental-health disorders are highly speaker-specific. Also, the speakers for depressed and healthy controls were different. Feature normalization for reducing speaker variabilities improved recognition performance when MFCC and formant-based features were used. All these techniques relied on hand-crafted features and traditional classifiers such

as GMM, SVM, etc., focusing on identifying relevant feature sets for robust classification of depressed speech from healthy controls.

Multi-modal approaches using audio, text, and facial geometry features have also been investigated in works such as [140–144]. The work in [140] investigated the fusion of information from speech, head pose, and eye gaze behaviors for depression/normal classification on a dataset of 30 depressed and 30 healthy controls collected by Black Dog Institute [145]. The central idea was to investigate using different feature selection and fusion techniques, and it is concluded that t-test-based feature selection performs well for binary depression/normal classification. Moreover, the individual modalities performance was also reported, with speech showing the maximum recognition accuracy of 83%, further strengthening the idea that speech alone contains sufficient information for robust depression detection. Also, in [141], new video and text features are proposed, and a hybrid of deep and shallow networks is used for depression classification using audio, video, and text modalities. Individual modalities such as audio and video were modeled using a DCNN-DNN-based system, while text modality was modeled using Paragraph Vector (PV) based SVM system. Moreover, in [143], an LSTM-based system was explored to simultaneously model depression from audio and text sequences without performing explicit topic modeling of the content of the interviews. Also addressing the AVEC 2016 depression sub challenge, the work in [144] used an i-vector framework with MFCC features for audio data modeling, and geometrical features along with polynomial parametrization of facial landmarks were used in a late-fusion fashion for depression classification. Different research works explored the different combinations of modalities to classify depression, and audio-based depression detection stands out from all other modalities, which motivates further exploring audio modality for robust depression classification.

With progress in the deep learning field and increased computation efficiency, the dependence on hand-crafted features is reduced. Much recent work has explored using time-frequency-based speech representations such as spectrograms and log-mel spectrograms as input for deep learning architectures to classify depression from audio. The work in [146] investigated spectrograms and raw waveforms as input to a CNN-based network on a subset of the DAIC-WOZ dataset in a speaker-dependent fashion. Moreover, in [147], a CNN-LSTM-based architecture was explored that extracted discriminative features from mel-spectrograms using 1d convolution in the first layer. A random sampling strategy was also proposed to mitigate the data imbalance associated with the DAIC-WOZ dataset. The majority voting of the labels for segments of speech coming from an individual is used

for depression prediction for an individual. In recent work in [148], an ensemble of 1d-CNN networks is used with mel-spectrograms as input features. The label for an individual is generated by the mean of the segment level probabilities for each constituent network in the ensemble. The ensemble labels are averaged to yield a final label for the individual. This ensemble technique showed appreciable improvements in recognition performance over hand-crafted features based on SVM classification and other single deep learning-based networks.

Multiple instance learning (MIL) is the apt choice when a single label is available for a group of utterances, as in Depression classification problem [149]. Most of the approaches in literature exploiting MIL architecture work by generating labels for individual segments and averaging them to yield a final label for the whole utterance. This is done using a network that shares parameters with all the segments of an utterance [150, 151]. However, the inherent problem with the MIL framework for depression classification is that not all the segments of the utterance exhibit depression-related characteristics, with the majority of the segments being in a neutral emotional state. As such, false labels are often predicted due to the majority of neutral state segments. Motivated by this issue, we propose a Tensor-based approach to extract shared and discriminative features from multiple segments of an utterance. Tensor factorizations provide a natural method for analyzing common information spread across modes of a tensor [44]. Utilizing this aspect, we use tensor factorization in conjunction with neural network-based learning to address the multiple-instance learning in a novel framework. Furthermore, the utterance level tensor core generated by the feature extraction block is passed on to an attention mechanism to generate the utterance level attentive feature. Statistic pooling of attentive representations is performed to extract bag-level features, which are classified using fully connected layers. This mitigates the dependence on average/max pooling output labels for individual segments for utterance level prediction, thus countering the inherent issue of traditional MIL frameworks.

6.2 Dataset and Preprocessing

For the task of depression classification from speech signals, we use the audio modality from the Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ), which is a subset of the larger corpus DAIC [152] and was introduced in the Audio/Visual Emotion Challenge (AVEC) 2016 [153]. The dataset consists of clinical interviews conducted between a participant and a virtual interviewer *ellie* which a human interviewer remotely controlled. The dataset was collected with the motive to

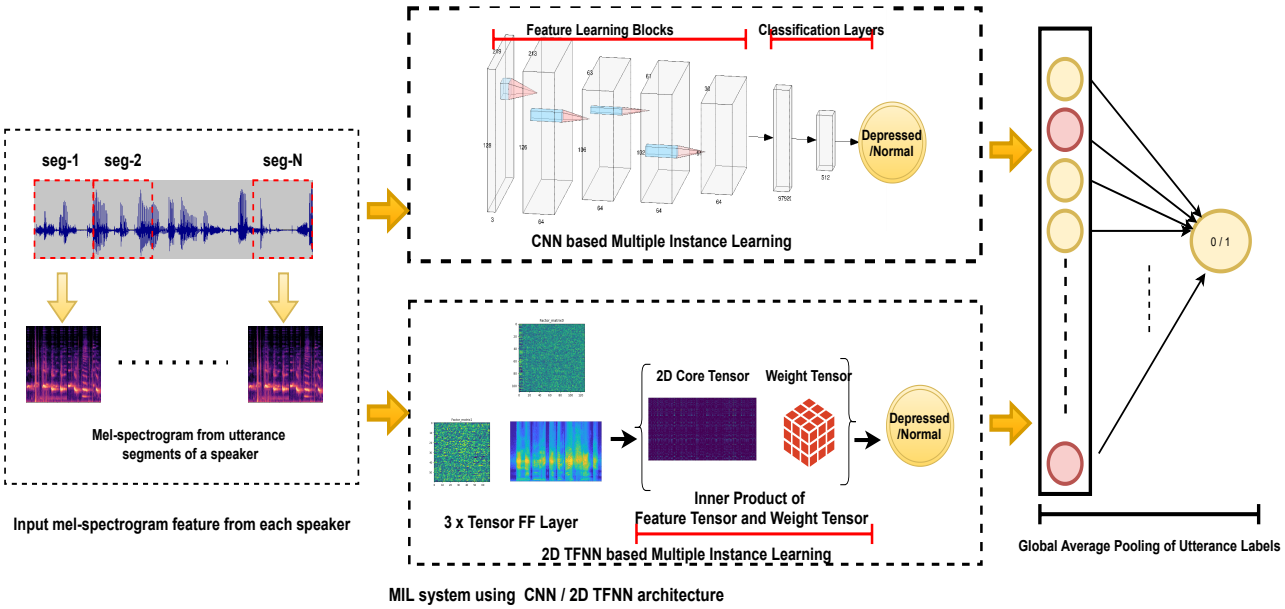


Figure 6.1: MIL Technique using CNN and TFNN as base architectures

augment the diagnoses of psychological conditions such as stress, anxiety, depression, etc., through automatic computer applications based on verbal and non-verbal indicators. It consists of audio, facial geometry features, and text transcriptions of the interviews. The dataset is recorded in English from a population of 189 subjects comprising 146 depressed subjects and 43 healthy controls. The audio duration ranges from 7-33 min (an average of 16 minutes). Each participant's audio file has been given a PHQ-8 score by the psychologist, which denotes the severity of depression, with 0 being no depression to 22 being severely depressed. Also, a binary PHQ-8 score is provided, classifying participants as depressed/not-depressed. Furthermore, the train-development-test split provided by the AVEC 2016 challenge divides the dataset into partitions comprising 118, 24, and 47 participants in the train, development, and test set, respectively.

Since the virtual interviewer's speech is not a part of the analysis, a silence region-based segmentation technique from the Python library *pyAudioAnalysis* [154] is employed to segment out the participant's speech and discard the speech segments from the virtual interviewer as it does not contain any emotion information. Also, the speech segments produced are of different duration, and deep learning techniques such as CNN and TFNN [80] require equal length input, so the speech segments are either zero-padded or truncated to seven seconds duration. The sampling rate of the speech signal is 16 kHz.

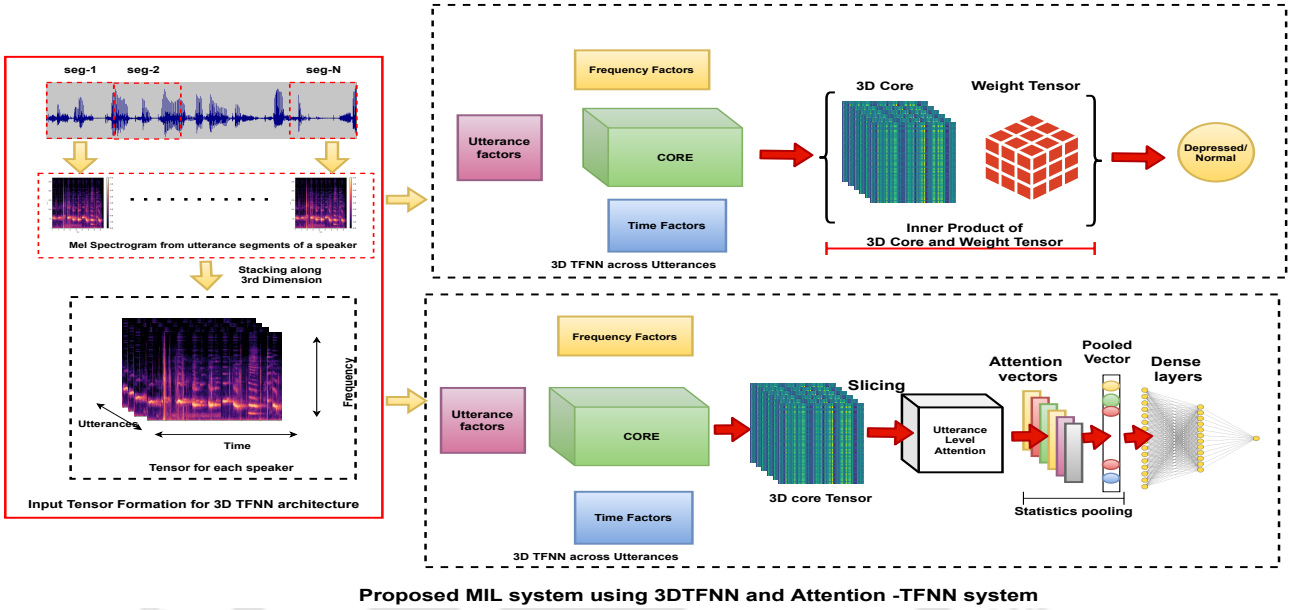


Figure 6.2: MIL Technique using 3D TFNN and 3D TFNN + Utterance level Attention as base architectures

6.3 Methodology

This section discusses the Tensor Factorization-based Multiple-Instance Learning Technique, which is used to classify depression versus normal speech from multiple utterances of a single speaker. Furthermore, an utterance level attention followed by a statistics pooling layer [155] is employed to extract temporal features in the subsequent layers of the network. Moreover, a standard Multiple-Instance Learning (MIL) network based on Convolution layers is also discussed, which serves as a baseline for comparing results.

6.3.1 CNN and 2D TFNN based MIL framework

Multiple Instance Learning with CNN as a base architecture has been explored in many previous works [156,157]. We have used this architecture as a baseline in our work. The base CNN architecture comprises 3 feature learning blocks followed by vectorization of the deep features and classification using a sigmoid layer. Each feature learning block comprises a 2D convolution layer, a batch normalization layer, an activation layer, and a max-pooling layer. The convolution layer extracts local features with the help of trainable kernels. Batch normalization forces the mean of the features over the entire batch to be centered at zero with unit variance. The normalized features are passed through an activation function (ELU in our work). Finally, a max-pooling layer is employed to reduce the size

of the feature maps obtained, keeping the relevant information only. Given a bag of utterances belonging to a speaker, the base CNN architecture is employed on each of the utterances to yield a label for each utterance. A global max-pooling of the labels yields the final label for the bag of utterances.

A 2D TFNN architecture [158] is employed as a base network for the MIL, similar to the CNN architecture. The 2D TFNN base receives mel spectrograms extracted from speech utterances as input. The factor matrices corresponding to the time and frequency modes extract the core feature tensor from the input tensors. Four consecutive Tensor FF layers yield the final feature tensor. This is then used to generate a class probability by doing an inner product with a weight matrix of the exact dimensions as the feature tensor.

6.3.2 3D TFNN Architecture as Feature extractor for MIL

The 3D TFNN architecture was introduced in [158] for emotion recognition from speech. The 3D TFNN serves as a natural framework for Multiple Instance Learning as the core idea of Tensor Factorization is capturing the shared information across different modes of a tensor. Given a bag of utterances belonging to a speaker, the utterances are first converted to the 2D speech representations such as mel-spectrograms of dimensions $I_{freq} \times I_{time}$. The mel-spectrograms for each utterance are stacked along the third dimension to form a 3D-tensor of dimensions $I_{freq} \times I_{time} \times I_{utter}$ representing the bag of utterances. The 3D tensor is passed through successive Tensor Factorization layers to obtain the deep feature tensors. Finally, a tensor sigmoid layer, comprising a weight tensor of the same size as the deep feature tensor, is utilized to get the probability for the bag of utterances.

The 3D TFNN architecture for Multiple Instance Learning benefits from not repeating the same architecture individually on each utterance as in conventional CNN-based MIL systems. Moreover, the probability generated by the 3D TFNN represents the entire bag as opposed to CNN-based MIL, where a global max-pooling of the labels generates a bag-level label. This comes from the inherent capability of Tensor Factorization-based feature extraction. The shared information across mel-spectrograms of utterances for an individual is utilized to conclude the label for that particular speaker. In contrast, the utterance level information is independent in conventional MIL systems, and no shared information across utterances is utilized.

6.3.3 3D TFNN with Utterance Level Attention

In this technique, the 3D TFNN described in 6.3.2 is utilized to extract deep tensor features from 3D tensor representations of bags of utterances. The feature tensor now comprises utterance-level representations stacked along the third dimension of the feature core tensor. For each 2D slice of the 3D feature tensor, an utterance level attentive feature representation is generated using the following attention mechanism.

6.3.3.1 Attention Layer

The attention layer used in our work is based on the attention proposed in [27]. The attention layer takes in a sequence of high-level feature vectors, focuses on the depression-related parts employing attention weights, and generates an utterance-level attention feature vector representing the depression-related frames of the input sequence. Given a 2D slice $\mathbf{H} \in \mathbf{R}^{I_2 \times I_3}$ of 3D feature tensor tensor $\mathcal{X} \in \mathbf{R}^{I_1 \times I_2 \times I_3}$, where I_1, I_2, I_3 represents the number of utterances, number of mel filter bands and number of frames respectively, normalized attention weights are first computed using a softmax function as described in equation -

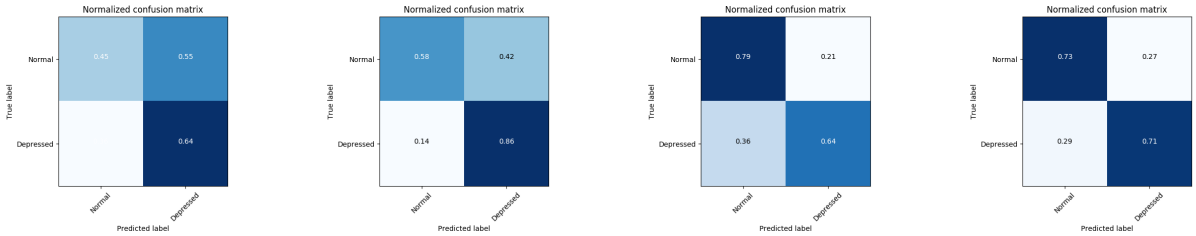
$$\alpha_t = \frac{\exp(W.h_t)}{\sum_{t=1}^{T} \exp(W.h_t)} \tag{6.1}$$

where $t \in (1, 2, \cdots, T)$, T being the total number of frames in the feature tensor slice and h_t being a feature vector belonging to the tth frame. The utterance level feature vector is obtained by taking the weighted sum of the attention weights with h_t as follows -

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t h_t \tag{6.2}$$

6.3.3.2 Statistics Pooling

The statistics pooling was first introduced in [155] for extracting utterance level statistics from frame-level features embeddings generated using a Time Delay Neural Network for speaker verification tasks. In our proposed architecture, statistics pooling is employed to extract bag level statistics - mean and standard deviation from the utterance level attentive feature vectors. As such, the output of the statistics pooling layer aggregates the relevant discriminative information obtained from several speaker utterances and provides a unified feature for further classification objectives. Given a set of



Baseline CNN-MIL TFNN-MIL 3D TFNN 3D TFNN+Attention

Figure 6.3: Normalized Confusion matrix for the test set of DAIC-WOZ Depression dataset for the three architectures - baseline CNN-MIL, TFNN-MIL,3D TFNN, and 3D TFNN+Attention.

attention feature vectors $C = (c_1, c_2, \cdots, c_{I_1})$ and $c \in \mathbf{R}^{I_2}$, obtained as described in section 6.3.3.1, where I_1 represents the number of utterances in the bag, the statistics pooling is calculated using *mean*, which is the average and *var*, which is the variance -

$$\mu = mean(C)$$

$$\sigma = var(C)$$

This results in a pooled feature vector of dimensions $\mathbf{R}^{2\times I_2}$, with μ and σ concatenated for each entry of c.

6.3.3.3 Fully Connected Layer

The output from the statistics pooling layer contains the aggregation of information across several utterances of a speaker. The pooled feature vector is passed to a fully connected network, having two layers to reduce the dimensionality and extract additional high-level features. Finally, the output of the fully connected layers is passed on to the last layer with sigmoid activation to generate the classification probability of being depressed/normal.

6.3.4 Experimental Setting

The four architectures - baseline CNN-MIL, TFNN-MIL, 3D TFNN, and 3D TFNN+Attention, are evaluated on the DAIC-WOZ dataset for Depression classification. For tensor formation, a set of utterances or bag sizes in the range $[10, 60]$ are selected from each speaker. Thus multiple tensors are formed for each speaker considering multiple bags formed because of the bag size chosen without repetition of utterances. For the training scenario, each bag of utterances is considered to come from a new speaker bearing the same label as all the other children bags of the parent speaker, thereby

generating a large number of tensors for training. However, for the testing scenario, the label for the parent speaker is calculated by averaging the predicted probability of all the children bags and comparing the final averaged probability against a threshold. The threshold is calculated from the ROC curve generated using the validation data.

Mel spectrograms are computed from the speech segments to be used as input for the Tensor Factorized Neural Network and baseline CNN architecture. For the computation of mel spectrograms, the speech segments are first windowed using a hamming window of 2048 samples with a shift of 512 samples. The windowed signal is used to compute Short-Time Fourier Transform (STFT). The magnitude spectrogram obtained from STFT is then passed through a mel-scale to obtain the filterbank energies. A log operator is finally used to get the log-mel spectrogram.

For baseline CNN architecture, the number of filters in the first and second feature learning block is 64 with a kernel size of 3×3 and a shift of 1. The number of filters for the third feature learning block is 128 with kernel size 2×2. The activation function used in all feature learning blocks is ELU and a max-pooling with a kernel size of 2×2 is used. The feature maps generated after the third feature learning block is vectorized and passed through a fully connected network with sigmoid non-linearity in its last layer to generate probabilities for the depressed versus non-depressed categories.

For the TFNN-MIL system, the base architecture consists of four consecutive 2D Tensor Feed Forward layers. The features dimension produced from the Tensor FF layers are respectively 120×210, 110×200, 100×180 and 80×160 . The output from the fourth Tensor FF layer is used to calculate logits using an inner product with a weight tensor of dimensions 80×160. Finally, the logits are passed through the activation function to yield utterance segment-level probabilities. This base architecture is repeated for all the instances in the bag, and a final global average pooling of the probabilities generates the bag level probability.

For 3D TFNN architecture, the input tensor is of size $num_{utter} \times 128 \times 219$ where the dimensions refer to the number of utterances, mel filters, and the number of time frames, respectively. The input mel-spectrogram tensor is passed through two 3D tensor feed-forward layers where the core tensors are of size $num_{utter} \times 120 \times 200$ and $num_{utter} \times 100 \times 180$ respectively. The activation function used in the Tensor FF layers is $RELU$. The feature tensor obtained after the second Tensor FF layer is fed to a Tensor sigmoid layer. The output of the inner product of the feature tensor with a trainable weight tensor of the same size is passed through a sigmoid non-linearity to generate class probability.

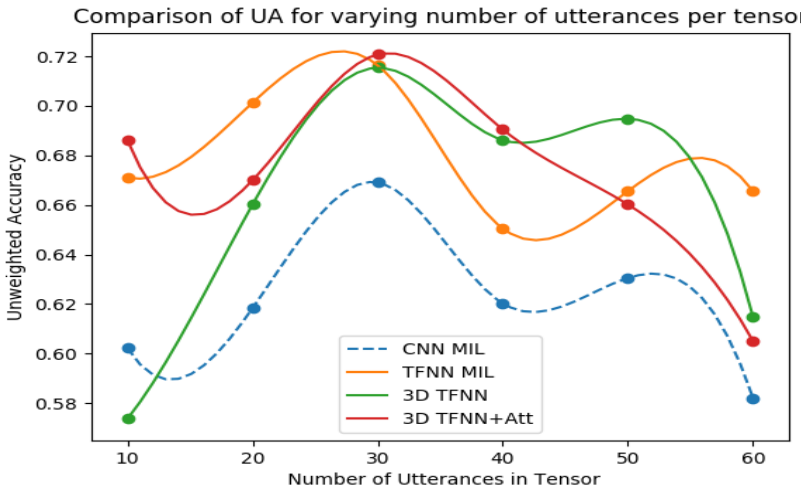


Figure 6.4: Comparison of Unweighted Accuracy for the varying number of utterances per tensor for the architectures CNN-MIL, TFNN-MIL, 3D TFNN, and 3D TFNN+Attention

Table 6.1: Recognition Accuracies in terms of Weighted Accuracy (WA) and Unweighted Accuracy(UA) and F1-scores for different Tensor Based Techniques for the test set of DAIC-WOZ Dataset

Method	Single Utterances		Speaker Level		
	WA	UA	WA	UA	F1-score (Normal, Depressed)
CNN MIL	54.40	55.65	51.06	54.87	0.56,0.43
TFNN MIL	60.00	62.52	65.95	71.64	0.70,0.60
3D TFNN	59.20	65.17	74.47	71.54	**0.81,0.60**
3D TFNN + Att	60.40	61.06	72.34	**72.07**	0.78,0.60

In the case of 3D TFNN+ Attention architecture, two 3D tensor FF layers, as used in 3D TFNN architecture above, extract discriminative feature tensor of the size $num_{utter} \times 100 \times 180$. The utterance level attention mechanism generates utterance level feature vectors of dimensions $num_{utter} \times 100$. This feature sequence is passed to a statistics pooling layer generating a feature vector of dimensions $\mathbf{R}^{200}$, which is passed through two fully connected layers of dimensions 256, 256 and a last layer having sigmoid non-linearity to generate a class probability for the bag of utterances.

6.4 Results

The four architectures - baseline CNN-MIL, TFNN-MIL, 3D TFNN, and 3D TFNN+Attention, are trained and evaluated on the DAIC-WOZ dataset using the following metrics - weighted accuracy, unweighted accuracy, and F1-score. Since the dataset is highly imbalanced, unweighted accuracy and F1-score become the apt choice to highlight the true prediction capability of the models. Moreover, another inherent issue with class imbalanced datasets is threshold-moving, which shifts the default threshold of 0.5 for binary classification problems. For our work, we have utilized the optimal threshold calculated from the ROC curve on the validation dataset, which is the development partition of the dataset. The optimal threshold is then used to generate labels for the probabilities predicted for the test set.

As seen from the table 6.1, the 3D TFNN and 3D TFNN + Attention architecture outperforms the baseline CNN-MIL system by a considerable margin of 16.67 percentage point and 17.2 percentage point respectively in terms of UA. This justifies that Tensor Factorized Neural Networks are more suitable for MIL-based systems due to their common information capturing capability amongst several modes of the tensor input. Moreover, the 3D TFNN+Attention system provides a balance of overall accuracy to an average of class accuracies. This becomes important for imbalanced datasets where the model's chances of fitting towards the majority class are always high. Moreover, in terms of F1-score, 3D TFNN outperforms other techniques and reaches the state-of-the-art.

Figure 6.3 presents the confusion matrices for the four architectures on the test set of the DAIC-WOZ dataset, taking 30 utterances per tensor. It is evident from the confusion matrix in Figure 3d that 3D TFNN+Attention architecture can balance the model toward both depressed and non-depressed categories, followed by 3D TFNN architecture. This supports our proposal of using utterance level attention to generate attentive feature vectors per utterance segment. Moreover, the impact of the number of utterances per tensor on the recognition performance of the model is assessed in Figure 6.4. The range of utterances per tensor is considered in the interval $[10, 60]$. The figure is plotted using b-spline interpolation [159] to account for the fewer data points and get a smooth curve. As is evident from the graph, the model performs best when 30 utterances are chosen per tensor. Also, the performance shows a gradual decline in the accuracy when the number of utterances per tensor is increased. This may be because redundant information apart from the desired objective is also being captured with increasing utterances, which accounts for increased confusion and decreased accuracy.

Table 6.2: Comparison with the state-of-the-art techniques on the test partition of DAIC-WOZ Dataset in terms of Weighted Accuracy(WA), Unweighted Accuracy(UA), and F1-scores.

sl.no	Method	Year of Publication	Accuracy		F1 score		
			WA	UA	Depressed	Normal	Mean
1.	Valstar et al (AVEC base)	2016	-	-	0.41	0.58	0.495
2.	Ma et al. (DepAudioNet)	2016	0.65	-	0.52	0.70	0.610
3.	Romero et al.(Ensemble)	2020	0.72	-	0.63	0.78	0.705
4	**3D TFNN (proposed)**	**-**	**0.745**	**0.715**	**0.60**	**0.81**	**0.705**

6.4.1 Comparison with State-of-the-Art

Several works have utilized DAIC-WOZ Depression dataset for unimodal and multi-modal depression recognition. Since in our proposed work, we have considered only the audio modality, therefore only those works or results are used for comparison which uses audio modality only. Moreover, few works have reported the final results on the development partition of the dataset. Our work utilizes the test set as the unseen data; we compare with similar works reporting results on test partition. Also, the works are segregated upon the metrics used to give a fair comparison. Only those works which have used accuracy and F1-score as metrics have been included for comparison.

Table 6.2 presents the state-of-the-art techniques for Depression recognition from speech utterances using the DAIC-WOZ dataset. The work in [153] provided the baseline results for the DAIC-WOZ dataset using both the audio and video modality. Our proposed technique outperforms the baseline by 0.21 for the mean F1 score for the audio modality scenario. Also, the work in [147] uses a combination of CNN and LSTM networks to extract high-level features from raw speech representations. It uses a random sampling strategy to balance the examples between depressed and normal classes. In contrast, our work uses a weighted loss function to alleviate the imbalance of classes and thereby incorporate all the training speakers during model training. As such, our proposed architecture achieves an overall performance gain of around 9 percentage point in terms of accuracy.

Conclusion

In this work, we proposed a tensor-based architecture for the task of Multiple Instance Learning when a bag of utterances for a speaker is available, and inferences about the speaker label must be drawn using the bag of utterances. The conventional MIL architectures, such as the baseline CNN-MIL system described in Figure 6.2 suffer from the inherent drawbacks of not considering relative and shared

information across the utterances in a bag. These techniques rely on inferring labels for individual utterances and averaging or max-pooling the labels to infer the speaker-level labels. The tensor-based architectures solve this problem by considering the utterances as the third mode in addition to the time and frequency modes in speech spectrograms. As such, TFNNs, by their rich mathematical framework, try to capture the shared information across the utterances of a bag by tensor factorization where the input tensor is projected over three subspaces - time subspace, frequency subspace, and utterance subspace. This helps to utilize the shared information and generate a single speaker/bag level probability for the specified task. To this end, we proposed two tensor MIL architectures - 3D TFNN and 3D TFNN+Attention. Comparison with the state-of-the-art proves that both the proposed techniques effectively capture depression-related information across bags of utterances. Moreover, additional analysis on the optimal number of utterances per bag is also presented to shed light on the model performance when using varying bag sizes.

7

Depression and Emotions : A Multimodal Approach

Objective *Since emotions in speech are highly affected due to an individual's underlying mental health issues, it becomes highly relevant when it comes to automatic assessment of clinical depression using speech. To this end, we propose emotion information fusion using Tensor-based fusion approaches. Two fusion approach is explored - Inner-Product based fusion and Elementwise Weighting. The emotion embedding tensors to be fused are generated using pre-trained TFNNs on six English SER datasets since the Depression dataset is also English. Moreover, a multi-modal and multi-task learning approach is also explored using audio and text modalities. BERT-based embeddings are explored for text transcripts and fused with audio embeddings learned from mel-spectrogram representations. Two fusion approaches are explored - Late feature fusion and Score Fusion.*

7.1 Introduction

The emotional state underlying an utterance can be related to Mental Health issues such as Depression, Anxiety, etc. [160]. Hall et al. [161] have shown that decreased verbal activity and monotonous and lifeless speech indicate depression. Moreover, according to Darby et al., there is a perceptible change in the pitch, speaking rate, loudness, and articulation of depressed patients before and after treatment [162]. Also, the work in [163] has shown improved recognition performance using emotional ratings in depression classification.

However, the incorporation of emotional information is not that straightforward. The major problem with this is the unavailability of a dataset with depression and emotion labels for utterances. Generally, depression labels are available on a speaker basis, while emotion labels are required at the utterance level as all the utterances of a speaker may not be emotionally monotonous. As such, the problem of incorporating emotional information in depression recognition becomes challenging.

7.1.1 Emotion Label Generation

The DAIC-WOZ Depression dataset used in our study contains binary labels of depressed/not-depressed for an individual as well as a depression severity score [153]. There are no labels for representing emotional information in the dataset mentioned above. This poses a challenge to identify appropriate emotion labels for the utterances of an individual to be used as ground truth for incorporating emotional information for aiding depression diagnosis. In literature, techniques such as unsupervised learning using clustering approaches and semi-supervised learning using a small amount of labeled data are explored to generate labels for unlabelled data [164]. Motivated by such ap-

Dataset	Total Utterances	WA	UA
SAVEE	301	68.65	66.45
RAVDESS	672	69.26	70.55
JL CORPUS	952	86.38	86.48
TESS	1600	100	100
ENTERFACE	846	82.94	84.54
IEMOCAP	5531	53.56	55.78

Table 7.1: Recognition performances and total utterances for the six emotion datasets of English language considered for emotion label generation

Classes	Angry	Happy	Neutral	Sad
Normal	624	461	3038	7630
Depressed	214	143	761	2262

Table 7.2: Distribution of emotion labels of the utterances from DAIC-WOZ Depression Dataset

proaches, we utilized semi-supervised learning to extract emotion labels for the utterance segments of the individuals.

To generate emotion labels for the individual utterances, we utilize emotion models trained on six Emotional speech datasets belonging to the English language so that there is no mismatch in language between the emotion and depression datasets. The models are trained for four emotion classes - Angry, Happy, Neutral, and Sadness using mel-spectrograms as input features and a 2D TFNN-based network described in Chapter 4. Given an utterance whose emotion label is to be predicted, the labels from the six pre-trained models are generated for the utterance. A mode operation, which counts the maximum occurrence of a label, is performed on the set of labels from the six pre-trained models to yield the final emotion label of the utterance. Table 7.2 shows the distribution of labels across the four emotion classes for utterances of the DAIC-WOZ Dataset.

7.1.2 Fusion of Emotion Information using speech modality

To incorporate emotional information using speech signals only, we proposed the Tensor embedding fusion techniques described below. The two fusion approaches utilize a set of pre-trained emotion networks which extracts feature tensors from speech mel-spectrograms. The feature tensors are stacked along the third dimension to form a third-order emotion embedding tensor. The network is trained for depression recognition, with added information from the emotion-sub network, non-trainable to

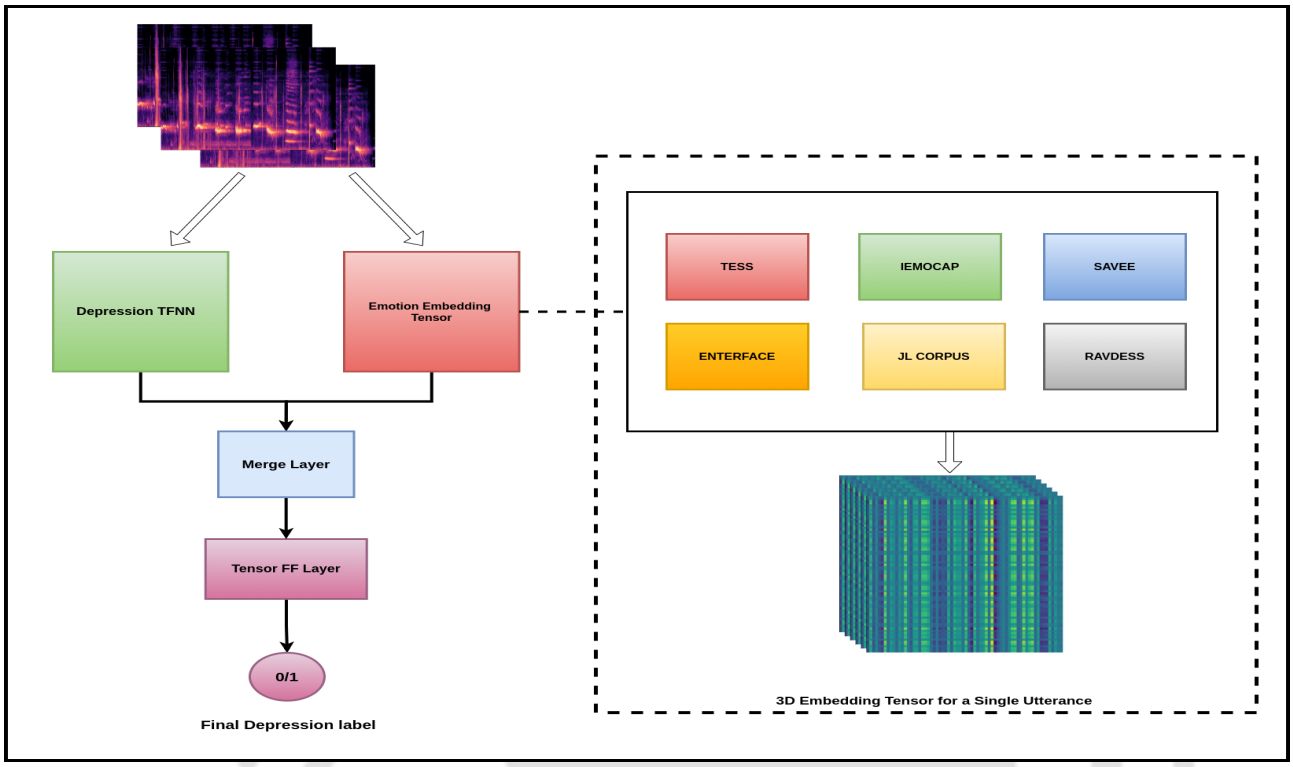


Figure 7.1: Multiple Instance Learning using conventional approach using a CNN architecture

extract emotion embedding tensor from speech utterances, with only the depression network having trainable parameters.

Proposed techniques for tensor embedding fusion are discussed below. Both the fusion techniques require mel-spectrograms as input to the two branches - the emotion embedding branch and the depression recognition branch.

(i) **Inner Product based Fusion**: This technique comes from the motivation of converting tensors to vectors via the inner-product of two equal-sized tensors as used in Chapter 4. The core tensor obtained from the depression sub-network and the emotion embedding sub-network is the same size and utilized for the inner product along the third dimension of the emotion embedding tensor, which contains model-specific embeddings stacked. Each of the six emotion pre-trained TFNN models generates 2D feature tensors. The six feature tensors are staked along the third mode to form a third-order tensor. This generates a feature vector with a length equal to the number of emotion models used for generating emotion embeddings. The feature tensor thus produced incorporates emotional information for subsequent classification by fully-connected layers. Mathematically, for an input mel-spectrogram tensor $\mathcal{X} \in \mathbf{R}^{I_{freq} \times I_{time}}$, where I_{freq}

represents the mel frequency bins and I_{time} represents the time information in terms of number of frames, the output of the depression TFNN sub-network will be -

$$\mathcal{A}_{dep} = TFNN(\mathcal{X}) \tag{7.1}$$

where TFNN represents the stack of Tensor FF Layers through which $\mathcal{X}$ is passed to obtain feature tensor $\mathcal{A}_{dep}$ of dimensions $J_{freq} \times J_{time}$ and $J_{freq} \leq I_{freq}$, $J_{time} \leq I_{time}$.

The emotion embedding sub-network takes in $\mathcal{X} \in \mathbf{R}^{I_{freq} \times I_{time}}$ as input and produces an embedding tensor $\mathcal{A}_{embed} \in \mathbf{R}^{J_{freq} \times J_{time} \times J_{model}}$, with the embeddings of each emotion model stacked along the third mode to form a third-order tensor with dimensions being J_{model}. The next step involves the inner product of the emotion embedding tensor $\mathcal{A}_{embed}$ with the depression feature tensor $\mathcal{A}_{dep}$ along the third mode, described mathematically as -

$$Feature_{fused} = \langle \mathcal{A}_{embed}, \mathcal{A}_{dep} \rangle \tag{7.2}$$

where $\langle a, b \rangle$ denotes inner-product between variables a and b of same size, as described in Chapter 2. The feature vector generated $Feature_{fused}$ is of dimensions $1 \times J_{model}$, which is passed to fully-connected layers followed by a sigmoid layer to generate class probability.

(ii) **Element wise Weighting**: This technique is motivated by the Tensor Attention Layer described in chapter 4. The core idea of this technique is to emotionally weight the elements of the feature tensor of depression sub-network and pass it to subsequent tensor layers for classification. To this end, we calculate the mean of the emotion embedding tensor along the third dimension, which serves as emotion weights for the core-tensor obtained from the depression sub-network. Hadamard product of the depression core-tensor and mean embedding tensor yields emotionally weighted feature tensor for classification by the subsequent layers. Mathematically, for an input mel-spectrogram tensor $\mathcal{X} \in \mathbf{R}^{I_{freq} \times I_{time}}$, the depression sub-network generates feature tensor $\mathcal{A}_{dep}$ as described in equation 7.1. Now, the mean of the emotion embedding tensor $\mathcal{A}_{embed} \in \mathbf{R}^{J_{freq} \times J_{time} \times J_{model}}$ is calculated along the third mode, which gives

$$\mathcal{A}_{mean} = mean(\mathcal{A}_{embed}) \tag{7.3}$$

Furthermore, the emotion weighting of feature tensor $\mathcal{A}_{dep}$ is performed by an element-wise

Sl no	Fusion Method	Accuracy		F1 score		
		WA	UA	N(D)	Macro avg	Wt. Avg
1	Inner Product	59.57	65.04	0.64(0.54)	0.59	0.61
2	Element wise multiplication with mean embed	65.95	67.53	0.72(0.56)	0.64	0.67
3	Method2 + Tensor FF	63.82	72.18	0.67(0.60)	0.64	0.65

Table 7.3: Recognition performances in terms of Accuracy and F1 scores for emotion information fusion-based techniques.

product, known as Hadamard product, as described below -

$$\mathcal{A}_{fused} = \mathcal{A}_{dep} \circ \mathcal{A}_{mean} \tag{7.4}$$

where $\circ$ represents Hadamard product. The emotionally weighted feature tensor $\mathcal{A}_{fused}$ is passed through subsequent Tensor FF Layers and Tensor Sigmoid Layer to generate class probability.

7.1.2.1 Experimental Evaluation

The fusion experiments are performed on DAIC-WOZ Depression dataset for depression modeling. The emotion embeddings are extracted using six English emotion datasets modelled using TFNN architecture- TESS [165], IEMOCAP [105], SAVEE [166], ENTERFACE [113], JL Corpus [167] and RAVDESS [168].

Table 7.3 shows the recognition performance for the two proposed fusion methods. It can be observed that Element-wise Weighting outperforms the inner-product-based fusion approach. However, using additional Tensor FF layers to model the fused feature tensor in the element-wise fusion method further improves recognition performance, thereby implicating efficient emotion incorporation, which aids the depression diagnosis system. Moreover, this technique (Method2 + Tensor FF) also ensures the balancing of F1 scores for the Normal and Depressed category compared to the other two approaches, which is a desirable outcome. The model reaches state-of-the-art performance in terms of Unweighted Accuracy (UA), indicating a balanced classification.

7.1.3 Text Based Emotion Information Fusion

Text transcripts contain significant sentiment information and have been used in numerous studies [169], [170]. Also, it becomes relevant in depression diagnosis as sentiments in a speaker's utterance

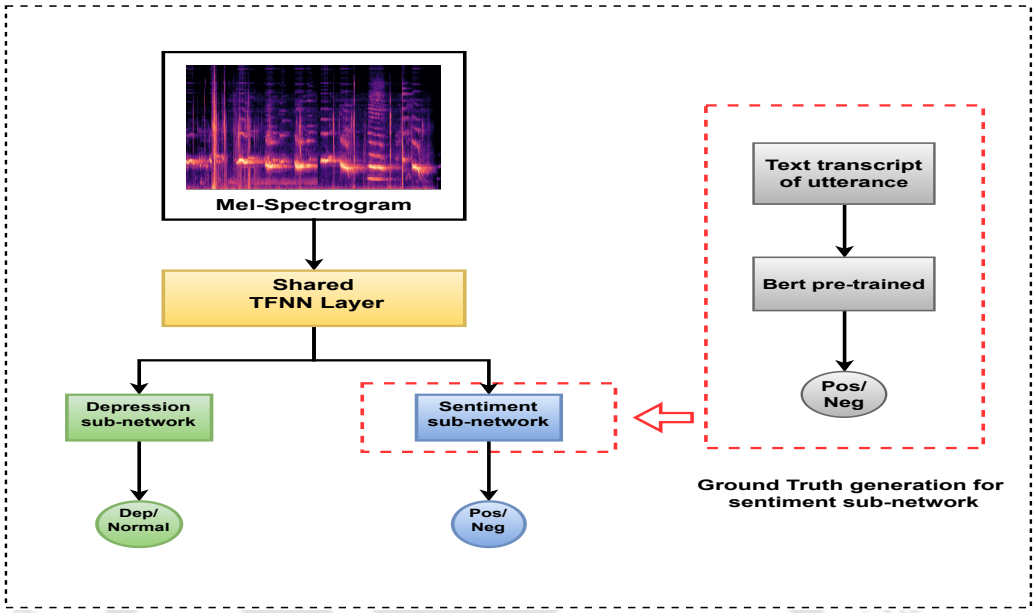


Figure 7.2: Multi-task Learning with depression and sentiment as two tasks. The sentiment labels are generated from text transcript using a pre-trained BERT sentiment analysis model.

Method	Accuracy		F1 -score		
	WA	UA	N(D)	Macro Avg	Wt. Avg
Text sentiment label	77.34	55.62	0.83(0.24)	0.53	0.65

Table 7.4: Recognition performance for multi-task architecture using depression and sentiment labels

can provide additional information that can help model depression-related aspects more efficiently. To this end, we explored two approaches - Multi-task Learning and Multi-modal Learning. Both these approaches use text as an additional modality alongside speech. A preliminary investigation of the proposed approaches is described below.

7.1.3.1 Multi-task Learning for Depression Diagnosis

Multi-task learning is a training technique where the same data is used as input, and the model is trained to learn multiple related tasks simultaneously, with some shared learning over the related tasks [171]. This has been shown to reduce computational load and data requirements and, at the same time, has contributed to improved modeling of target task [172]. Motivated by this, we proposed sentiment classification from the speech sub-network and the depression sub-network in a multi-task learning scenario. This method generates sentiment labels for each speech utterance using its corresponding

Method	Accuracy	
	WA	UA
Feature Fusion	78.72	72.51
Score fusion	63.82	61.90

Table 7.5: Recognition performance for multi-modal architecture using audio and text modality

text transcript. The text labels are generated using a pre-trained BERT-based sentiment classification model [173]. Each speech utterance now contains two label information - one depression label (N/D) and one sentiment label (Positive/Negative). As seen from Figure 7.2, a Tensor FF layer acts as the shared layer between two task sub-architectures. This shared layer tries to capture correlated information across the two tasks. An experimental evaluation of the described approach is shown in Table 7.4. It can be seen that adding sentiment information improves Weighted Accuracy (WA). However, the F1 scores for normal/depressed class are not balanced.

7.1.3.2 Multi-Modal Learning using Text information

The perception of paralinguistic phenomenon not only depends on speech signal alone but also on text transcripts, i.e., what is being said, facial expressions captured through images and videos, etc. [174]. Multi-modal information becomes significant when modeling complex phenomena such as clinical depression. Multi-modal fusion of information from modalities such as audio, video, and text has been investigated in studies [175] for depression recognition. In our work, we are considering only the audio and text modality as video modality for depression and other mental health datasets is not available due to ethical constraints.

For the text-based depression recognition system, we have utilized the BERT pre-trained sentiment analysis model as described in 7.1.3.1. The base architecture for modeling audio modality in the MIL framework is chosen to be CNN. This is done because BERT embeddings for sentences are vectors, and the CNN-based audio MIL branch will produce vector embeddings, which will be fused later.

Multi-modal fusion is considered in two ways -

(i) **Feature level fusion**: The BERT-based text features are fused with speech-based features in the MIL framework. A further classification network is trained using this fused feature.

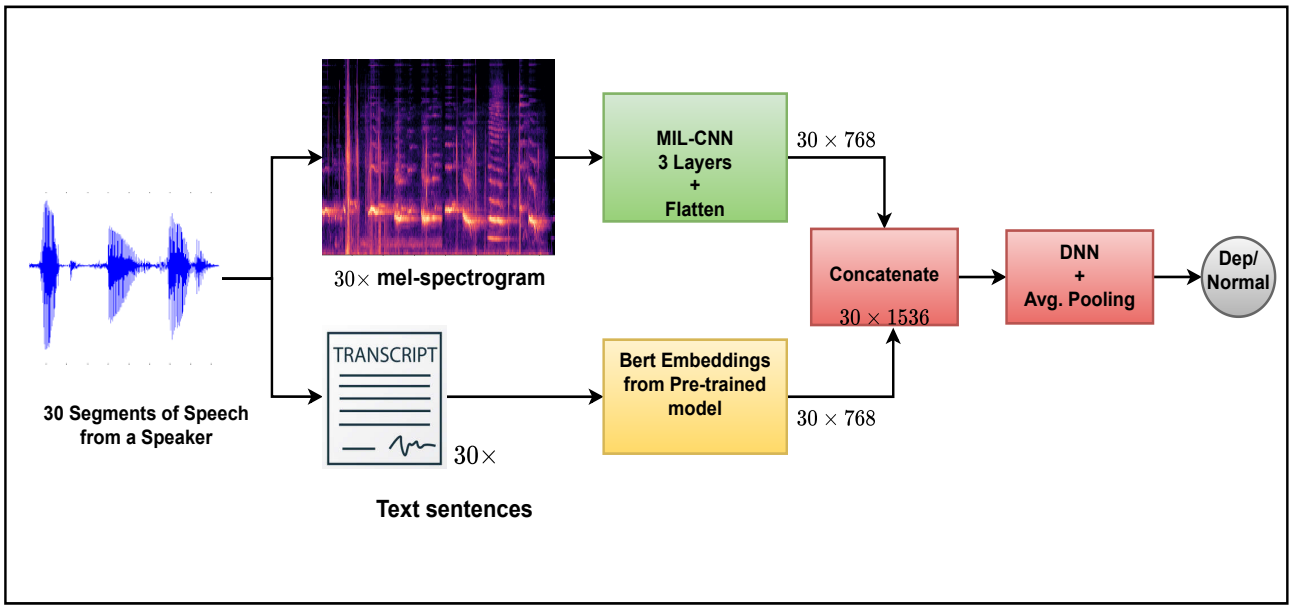


Figure 7.3: Feature fusion from audio and text modality

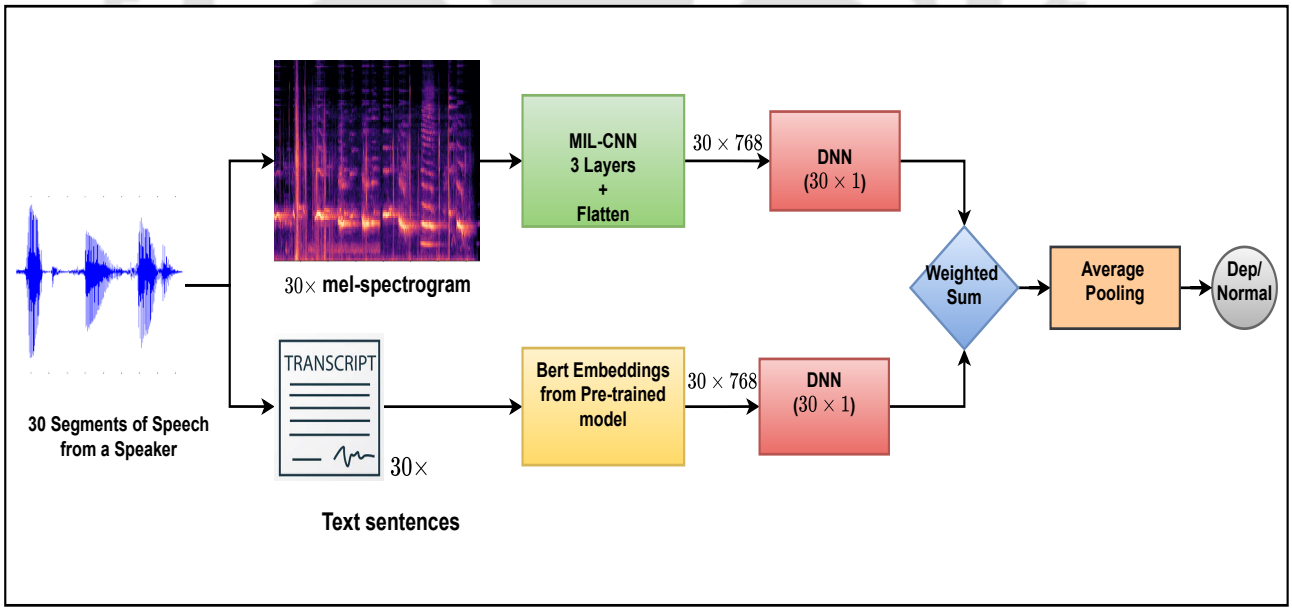


Figure 7.4: Score-level Weighted fusion from audio and text modality

Figure 7.5: Proposed Multi-Modal depression recognition system using audio and text modality.

(ii) **Decision level fusion**: Probability scores are generated using both text and audio modality for utterance levels. A weighted average probability score yields the final probability for the bag of utterances.

Experimental Evaluation of the proposed multi-modal framework using the two fusion approach - Feature fusion and Decision Level Fusion is demonstrated in terms of recognition accuracy in Table 7.5. From the table, it can be seen that Feature Level Fusion outperforms Decision Level Fusion and reaches state-of-the-art accuracy for the DAIC-WOZ dataset. This can be attributed to more abstract information learning from combined feature vectors in the last layers of the multi-modal framework.

7.1.4 Conclusion

This chapter proposes a combined framework for efficient fusion of emotional information to aid depression diagnosis from speech signals. To this end, emotion labels and embeddings are extracted for depression dataset utterances in a semi-supervised method. Six emotion datasets belonging to the English language are exploited for this purpose. Furthermore, the emotion embeddings are fused with depression embeddings using tensor fusion strategies, which accounts for efficient and intuitive fusion in tensor data. Moreover, to exploit the depression-related information present in the text, we explore multi-task and multi-modal architectures with BERT embeddings as text features. In the multi-task scenario, a BERT-based sentiment classification system generates additional sentiment labels for speech utterances corresponding to the text transcript. In the multi-modal scenario, feature level fusion and decision level fusion is explored in the MIL-CNN setting to get a first-hand feel of multi-modal depression diagnosis systems.

8

Summary and Conclusions

Objective *This chapter provides the summary and conclusions of the works presented in this thesis towards improving Human-Computer Interaction by enhancing the performance of speech emotion recognition using Tensor architectures and thereby use it to aid the automatic diagnosis of Mental health ailments such as depression and Alzheimer's dementia. Based on the contributions and different investigations, we also discuss a few possible directions for future research.*

8.1 Summary of the work

The present thesis addresses the problem of Speech Emotion Recognition (SER) from a tensor factorization perspective. The primary objective of this thesis is to present a solution in the form of Tensor Factorized Neural Networks (TFNN) to the inherent problems associated with conventional CNN+LSTM-based deep learning architectures such as high computational complexity, complex architecture, vectorization of features, etc. The TFNN efficiently handles such issues with improved recognition performance in terms of weighted and unweighted accuracy. Apart from the issues related to architecture, two more fundamental issues pertaining to SER, in general, have been explored in this thesis - dependence of emotions on the speaker and cultural specificity of emotions. These issues are core when practical applications of SER in Human-Computer Interaction (HCI) are considered, such as deploying voice assistants in the out-of-domain language area and mental health analysis where models are trained on a specific population but have to be used for unseen patients, etc. Two methodologies in this thesis have been proposed to tackle these two fundamental issues related to SER. Firstly, to incorporate speaker and language cues along with emotional cues to adapt to the speaker and language scenario, and secondly, to normalize the speaker and language influence to present a generalized model independent of such variabilities. Moreover, automatic diagnoses of mental health issues, a crucial aspect of SER, are explored both from SER and Tensor perspectives. The proposed Tensor-MIL method surpasses the limitations of the Conventional Multiple Instance Learning frameworks based on CNN as base architectures. The contributions incorporated in this thesis are summarized below -

(i) **Attention-Gated Tensor Neural Network architectures for SER :** The computational complexity problem, large parameter size, huge data requirement, vectorization of features, etc., are inherent to the conventional CNN-LSTM-based SER methods explored in the literature. Tensor Factorization based neural network architectures are proposed to tackle these problems, improving recognition performance while keeping the architecture simple and with fewer pa-

rameters. Moreover, it is shown that additional information in the form of third-order tensors with a 3D TFNN network performs better than the 2D counterpart, thereby justifying exploring features in tensor form rather than vector form. A parallel variant of TFNN is also explored, which can exploit complementary information from two different features.

(ii) **Robust and generalized SER by addressing Speaker and Language Dependence of Emotions:** The fundamental issue of language and speaker dependence is addressed. Two approaches are explored - one takes in additional speaker and language embeddings alongside emotion embeddings, and the other uses triplet loss-based TFNN architecture to normalize these variabilities.

(iii) **Depression recognition from a Tensor Perspective:** The inherent issue of not capturing the shared information from multiple instances of a speaker's utterance is addressed using Tensor-based MIL approaches. The tensor MIL approach, alongside utterance-level attention and statistics pooling, is proposed to tackle the MIL problem in depression recognition.

(iv) **Leveraging emotion cues and multi-modal information for Depression recognition:** Mental health of affected individuals highly correlates with emotions in their speech. This is explored using a Tensor-based fusion of emotional information to aid depression diagnosis. Furthermore, multi-task and multi-modal architectures are also explored to exploit the capabilities of text-based sentiment analysis, and a preliminary investigation of the same is also presented.

8.2 Contribution of the thesis :

Following are the contribution of the thesis toward efficient and robust SER using a tensor framework-

- Tensor factorization-based neural network architectures are proposed as an alternative to conventional CNN+LSTM-based architectures.
- A tensor attention layer is proposed, responsible for providing emotion salient attentive inputs to the network.
- 3D-AGTFNN is proposed to deal with tensor inputs of order-3. A third-order tensor is constructed using delta and double-delta information of mel spectrogram. The information presented in higher-order tensor form is more efficient than its lower-order counterparts.

- Parallel AG-TFNN is proposed to leverage complementary information from mel spectrograms and modulation spectrograms and fuse them in tensor form.
- To incorporate speaker and language dependence in a multi-lingual scenario, a fusion architecture is proposed where language and speaker embeddings are fused with emotion embeddings for further classification.
- In an attempt to generalize SER over cultures and speakers, a triplet-loss-based architecture is proposed, which has demonstrated a robust cultural adaptivity when tested in a multi-lingual scenario.
- To solve the inherent weak labeling problem of conventional MIL-based systems, a Tensor-based MIL is proposed and applied to Depression recognition.
- A novel attention pooling-based Tensor architecture is proposed, which extracts attentive features from multiple utterances in a tensor.
- Tensor-based Fusion strategies are explored to incorporate emotional information for robust depression classification.
- Text sentiment-based approach is explored to aid depression diagnosis from speech.

8.3 Directions for future work

Based on the outcome of this thesis work, this section provides some of the possible future work directions -

(i) Tensors are powerful mathematical tools for multi-dimensional data. Tensor construction is crucial as the trained model's quality is highly dependent on what information the input tensor holds. However, in our work, the tensor construction has been limited to audio modality and its 2D representations such as Mel Spectrograms and Modulation Spectrograms and the use of delta features in the case of 3D Mel spectrogram. As such, more efficient tensors can be explored, such as multi-scale Tensors using signal decomposition techniques like EMD, VMD, etc., and forming 3D tensors, with the third mode being the scale dimension.

(ii) Emotion information is known to be incorporated in facial expressions, voice, gestures, and text transcript of individuals. The work involving all such modalities usually relies on feature-fusion

or decision fusion of feature vectors from individual modalities. Tensors provide a natural fusion framework for multi-modal data; hence, the performance of emotion recognition systems can be improved using tensors without additional parameter load.

(iii) From chapter 4, it can be seen that the language model used for Method-1 is fixed-set, i.e., it contains the language present in the test set as well. However, when a new unseen language is presented, the embeddings produced using such a model will not be discriminative. This problem has to be addressed by building a universal language model incorporating as many languages as data and computational resources permit or fine-tuning the existing model with a small subset of the target language each time an unseen language is presented.

(iv) The methodologies proposed in chapter 4 has been tested using five languages - English, Hindi, Telugu, German and Persian. Based on the language similarity analysis in Table II, it is seen that some of the languages are genetically similar to some others, and as such, giving an implication of cultural similarity of emotions. As such, this multi-cultural study of emotions needs to be extended to the collection of datasets having a fair genetic dissimilarity within them.

Bibliography

[1] J. W. Pennebaker, *Emotion, disclosure, & health.* American Psychological Association, 1995.

[2] S. Yoon and J. Rottenberg, "Why do people with depression use faulty emotion regulation strategies?" *Emotion Review*, vol. 12, no. 2, pp. 118–128, 2020.

[3] L. Jiang, P. Tan, J. Yang, X. Liu, and C. Wang, "Speech emotion recognition using emotion perception spectral feature," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 11, p. e5427, 2021.

[4] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[5] H. Smith and A. Schneider, "Critiquing models of emotions," *Sociological Methods & Research*, vol. 37, no. 4, pp. 560–589, 2009.

[6] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[7] Z. Dair, R. Donovan, and R. O'Reilly, "Linguistic and gender variation in speech emotion recognition using spectral features," *arXiv preprint arXiv:2112.09596*, 2021.

[8] M. Sidorov, A. Schmitt, E. Semenkin, and W. Minker, "Could speaker, gender or age awareness be beneficial in speech-based emotion recognition?" in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 61–68.

[9] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[10] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden markov models," in *Proceedings of the 7th European Conference on Speech Communication and Technology*, 2001.

[11] H. Hu, M.-X. Xu, and W. Wu, "Gmm supervector based svm with spectral features for speech emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, vol. 4, 2007, pp. IV–413.

[12] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009.

[13] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013.

[14] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), INTERSPEECH*, 2017, pp. 3442–3446.

[15] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[16] J. H. Jeon, R. Xia, and Y. Liu, "Sentence level emotion recognition based on decisions from subsentence segments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4940–4943.

[17] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, vol. 4, 2007, pp. IV–17.

[18] R. Sun and E. Moore_II, "A preliminary study on cross-databases emotion recognition using the glottal features in speech," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association*, 2012.

[19] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[20] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7234–7238.

[21] O. Vinyals and G. Friedland, "Modulation spectrogram features for improved speaker diarization," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[22] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 836–845, 2014.

[23] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Proceedings of INTERSPEECH*, 2017, pp. 1089–1093.

[24] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proceedings of the International Conference on Platform Technology and Service (PlatCon)*, 2017, pp. 1–5.

[25] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition." in *Proceedings of INTERSPEECH*, 2016, pp. 3603–3607.

[26] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[27] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[28] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97 515–97 525, 2019.

[29] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using dnns." in *Proceedings of INTERSPEECH*, 2018, pp. 3097–3101.

[30] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5200–5204.

[31] S. K. Pandey, H. Shekhawat, and S. Prasanna, "Emotion recognition from raw speech using wavenet," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, 2019, pp. 1292–1297.

[32] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[33] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 801–804.

[34] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227–2231.

[35] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *CoRR*, vol. abs/1704.08619, 2017. [Online]. Available: http://arxiv.org/abs/1704.08619

[36] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90 368–90 377, 2019.

[37] C. Breitenstein, D. V. Lancker, and I. Daum, "The contribution of speech rate and pitch variation to the perception of vocal emotions in a german and an american sample," *Cognition & Emotion*, vol. 15, no. 1, pp. 57–79, 2001.

[38] T. Levi-Civita, *The absolute differential calculus (calculus of tensors)*. Courier Corporation, 1977.

[39] Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer, "Novel methods for multilinear data completion and de-noising based on tensor-SVD," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3842–3849.

[40] A. H. Phan and A. Cichocki, "Tensor decompositions for feature extraction and classification of high dimensional datasets," *Nonlinear theory and its applications, IEICE*, vol. 1, no. 1, pp. 37–68, 2010.

[41] L. De Lathauwer, B. De Moor, and J. Vandewalle, "Blind source separation by higher-order singular value decomposition," in *Proceedings of the EUSIPCO*, vol. 1, 1994, pp. 175–178.

[42] X. Guo, X. Huang, L. Zhang, and L. Zhang, "Hyperspectral image noise reduction based on rank-1 tensor decomposition," *ISPRS journal of photogrammetry and remote sensing*, vol. 83, pp. 50–63, 2013.

[43] A. Anandkumar, P. Jain, Y. Shi, and U. N. Niranjan, "Tensor vs. matrix methods: Robust tensor decomposition under block sparse perturbations," in *Artificial Intelligence and Statistics*. PMLR, 2016, pp. 268–276.

[44] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[45] H. S. Shekhawat and S. Weiland, "A novel computational scheme for low multi-linear rank approximations of tensors," in *Proceedings of the European Control Conference (ECC)*, 2015, pp. 3003–3008.

[46] A.-H. Phan, P. Tichavskỳ, and A. Cichocki, "Candecomp/parafac decomposition of high-order tensors through tensor reshaping," *IEEE Transactions on Signal Processing*, vol. 61, no. 19, pp. 4847–4860, 2013.

[47] L. R. Tucker, "Implications of factor analysis of three-way matrices for measurement of change," *Problems in measuring change*, vol. 15, no. 122-137, p. 3, 1963.

[48] Y.-D. Kim and S. Choi, "Nonnegative tucker decomposition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2007, pp. 1–8.

[49] S. Zubair and W. Wang, "Tensor dictionary learning with sparse tucker decomposition," in *Proceedings of the 18th International Conference on Digital Signal Processing (DSP)*, 2013, pp. 1–6.

[50] S. Y. Chang and H.-C. Wu, "Tensor wiener filter," *IEEE Transactions on Signal Processing*, vol. 70, pp. 410–422, 2022.

[51] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *Proceedings of the European conference on computer vision*. Springer, 2002, pp. 447–460.

[52] F. Cuzzolin, M. Sapienza, and W. Gong, "Fisher tensor decomposition for unconstrained gait recognition," 2013.

[53] D. G. Chachlakis, M. Dhanaraj, A. Prater-Bennette, and P. P. Markopoulos, "Options for multimodal classification based on l1-tucker decomposition," in *Big Data: Learning, Analytics, and Applications*, vol. 10989. International Society for Optics and Photonics, 2019.

[54] T. Lin and S. Bourennane, "Survey of hyperspectral image denoising methods based on tensor decompositions," *EURASIP journal on Advances in Signal Processing*, vol. 2013, no. 1, pp. 1–11, 2013.

[55] Y.-X. Wang, L.-Y. Gui, and Y.-J. Zhang, "Neighborhood preserving non-negative tensor factorization for image representation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 3389–3392.

[56] D. Yu, L. Deng, and F. Seide, "The deep tensor neural network with applications to large vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 388–396, 2012.

[57] Y. Shan, M. Liu, Q. Zhan, S. Du, J. Wang, and X. Xie, "Speech recognition based on deep tensor neural network and multifactor feature," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 650–654.

[58] C. Jutten and J. Herault, "Blind separation of sources, part-1: An adaptive algorithm based on neuromimetic architecture," *Signal processing*, vol. 24, no. 1, pp. 1–10, 1991.

[59] B. Makkiabadi, F. Ghaderi, and S. Sanei, "A new tensor factorization approach for convolutive blind source separation in time domain," in *Proceedings of the 18th European Signal Processing Conference*, 2010, pp. 900–904.

[60] T. Barker and T. Virtanen, "Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation." in *Proceedings of INTERSPEECH*, vol. 2813, 2013, pp. 827–831.

[61] Y. Xie, K. Xie, J. Yang, and S. Xie, "Underdetermined blind source separation combining tensor decomposition and nonnegative matrix factorization," *Symmetry*, vol. 10, no. 10, p. 521, 2018.

[62] J. Qi, H. Hu, Y. Wang, C.-H. H. Yang, S. M. Siniscalchi, and C.-H. Lee, "Exploring deep hybrid tensor-to-vector network architectures for regression based speech enhancement," *arXiv preprint arXiv:2007.13024*, 2020.

[63] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, 2011.

[64] S. Samui, I. Chakrabarti, and S. K. Ghosh, "Tensor-train long short-term memory for monaural speech enhancement," *arXiv preprint arXiv:1812.10095*, 2018.

[65] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[66] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[67] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[68] K. S. Rao and S. Sarkar, *Robust speaker recognition in noisy environments*. Springer, 2014.

[69] N. Dehak, "Discriminative and generative approaches for long-and short-term speaker characteristics modeling: application to speaker verification," Ph.D. dissertation, École de technologie supérieure, 2009.

[70] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[71] Y. Jeong, "Speaker adaptation based on the multilinear decomposition of training speaker models," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4870–4873.

[72] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[73] N. M. I. Group, "Nist speaker recognition evaluation ldc2010s03," 2010. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2010S03

[74] A. Krizhevsky, G. Hinton *et al.*, "Factored 3-way restricted boltzmann machines for modeling natural images," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 621–628.

[75] G. E. Hinton *et al.*, "Modeling pixel means and covariances using factorized third-order boltzmann machines," 2010.

[76] T. D. Nguyen, T. Tran, D. Phung, and S. Venkatesh, "Tensor-variate restricted boltzmann machines," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

[77] D. Yu, L. Deng, and F. Seide, "The deep tensor neural network with applications to large vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 388–396, Feb 2013.

[78] J. Wu, S. Qiu, R. Zeng, Y. Kong, L. Senhadji, and H. Shu, "Multilinear principal component analysis network for tensor object classification," *IEEE Access*, vol. 5, pp. 3322–3331, 2017.

[79] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?" *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

[80] J.-T. Chien and Y.-T. Bao, "Tensor-factorized neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1998–2011, 2017.

[81] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding." in *Proceedings of INTERSPEECH*, 2018, pp. 3688–3692.

[82] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2015, pp. 448–456.

[83] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[84] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125 868–125 881, 2019.

[85] L. R. Rabiner, R. W. Schafer *et al.*, "Introduction to digital speech processing," pp. 1–194, 2007.

[86] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–25.

[87] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, 2012, pp. 2951–2959.

[88] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[89] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[90] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.

[91] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[92] N. Semwal, A. Kumar, and S. Narayanan, "Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models," in *Proceedings of the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 2017, pp. 1–6.

[93] L. Guo, L. Wang, J. Dang, Z. Liu, and H. Guan, "Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine," *IEEE Access*, vol. 7, pp. 75 798–75 809, 2019.

[94] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using gans for speech emotion recognition." in *Proceedings of INTERSPEECH*, 2019, pp. 171–175.

[95] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH), Vols 1-5*, 2016, pp. 2001–2005.

[96] S. Ikeda, M. Sudo, T. Matsui, and E. Haryu, "Developmental changes in understanding emotion in speech in children in japan and the united states," *Cognitive Development*, vol. 60, p. 101110, 2021.

[97] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, and M. Shimura, "Usage of emotion recognition in military health care," in *Proceedings of the Defense Science Research Conference and Expo (DSR)*, 2011, pp. 1–5.

[98] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proceedings of artificial neural networks in engineering*, vol. 710, 1999, p. 22.

[99] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis." *Psychological bulletin*, vol. 128, no. 2, p. 203, 2002.

[100] M. T. Riviello, A. Esposito, and K. Vicsi, "A cross-cultural study on the perception of emotions: How hungarian subjects evaluate american and italian emotional expressions," in *Cognitive behavioural systems*. Springer, 2012, pp. 424–433.

[101] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[102] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[103] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based lstm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[104] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, 2005.

[105] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[106] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 585–589, 2016.

[107] E. M. Albornoz and D. H. Milone, "Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 43–53, 2015.

[108] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," *IEEE Annals of the History of Computing*, no. 02, pp. 265–275, 2019.

[109] A. Shaukat and K. Chen, "Exploring language-independent emotional acoustic features via feature selection," *arXiv preprint arXiv:1009.0117*, 2010.

[110] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *Proceedings of the International Conference on Frontiers of Information Technology (FIT)*, 2018, pp. 88–93.

[111] M. Neumann *et al.*, "Cross-lingual and multilingual speech emotion recognition on english and french," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5769–5773.

[112] S. Goel and H. Beigi, "Cross lingual cross corpus speech emotion recognition," *arXiv preprint arXiv:2003.07996*, 2020.

[113] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*, 2006, pp. 8–8.

[114] S. G. Koolagudi, R. Reddy, J. Yadav, and K. S. Rao, "Iitkgp-sehsc: Hindi speech corpus for emotion analysis," in *Proceedings of the International Conference on Devices and Communications (ICDeCom)*. IEEE, 2011, pp. 1–5.

[115] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, "Iitkgp-sesc: Speech database for emotion analysis," in *Proceedings of the International Conference on Contemporary Computing*. Springer, 2009, pp. 485–492.

[116] O. Mohamad Nezami, P. Jamshid Lou, and M. Karami, "Shemo: A large-scale validated database for persian speech emotion detection," *Language Resources and Evaluation*, vol. 53, no. 1, pp. 1–16, 2019.

[117] V. Beaufils and J. Tomin, "Stochastic approach to worldwide language classification: the signals and the noise towards long-range exploration," Oct 2020. [Online]. Available: osf.io/preprints/socarxiv/5swba

[118] S. K. Pandey, H. S. Shekhawat, and S. Prasanna, "Attention gated tensor neural network architectures for speech emotion recognition," *Biomedical Signal Processing and Control*, vol. 71, p. 103173, 2022.

[119] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[120] Z. Lian, Y. Li, J. Tao, and J. Huang, "Speech emotion recognition via contrastive loss under siamese networks," in *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, 2018, pp. 21–26.

[121] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1633–1644, 2018.

[122] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[123] J. Harvill, M. AbdelWahab, R. Lotfian, and C. Busso, "Retrieving speech samples with similar emotional content using a triplet loss function," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7400–7404.

[124] J. Huang, Y. Li, J. Tao, Z. Lian *et al.*, "Speech emotion recognition from variable-length inputs with triplet loss function." in *Proceedings of INTERSPEECH*, 2018, pp. 3673–3677.

[125] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[126] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS medicine*, vol. 3, no. 11, p. e442, 2006.

[127] K. Hawton, C. C. i Comabella, C. Haw, and K. Saunders, "Risk factors for suicide in individuals with depression: a systematic review," *Journal of affective disorders*, vol. 147, no. 1-3, pp. 17–28, 2013.

[128] M. Hamilton, "The hamilton rating scale for depression," in *Assessment of depression*. Springer, 1986, pp. 143–152.

[129] K. Kroenke and R. L. Spitzer, "The phq-9: a new depression diagnostic and severity measure," 2002.

[130] P. J. Cowen and M. Browning, "What has serotonin to do with depression?" *World Psychiatry*, vol. 14, no. 2, p. 158, 2015.

[131] P. E. Croarkin, A. J. Levinson, and Z. J. Daskalakis, "Evidence for gabaergic inhibitory deficits in major depressive disorder," *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 3, pp. 818–825, 2011.

[132] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Eye movement analysis for depression detection," in *Proceedings of the IEEE International Conference on Image Processing*, 2013, pp. 4220–4224.

[133] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–7.

[134] M. JH Balsters, E. J Krahmer, M. GJ Swerts, and A. JJM Vingerhoets, "Verbal and nonverbal correlates for depression: a review," *Current Psychiatry Reviews*, vol. 8, no. 3, pp. 227–234, 2012.

[135] C. Segrin, "Social skills deficits associated with depression," *Clinical psychology review*, vol. 20, no. 3, pp. 379–403, 2000.

[136] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, G. Parker, M. Breakspear *et al.*, "Characterising depressed speech for classification," in *Proceedings of INTERSPEECH*, 2013.

[137] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, G. Parker *et al.*, "From joyous to clinically depressed: Mood detection using spontaneous speech." in *Proceedings of the FLAIRS Conference*, vol. 19. Citeseer, 2012.

[138] H. Long, Z. Guo, X. Wu, B. Hu, Z. Liu, and H. Cai, "Detecting depression in speech: Comparison and combination between different speech types," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 1052–1058.

[139] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, 2011.

[140] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 478–490, 2018.

[141] L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 239–253, 2021.

[142] L. Yang, D. Jiang, W. Han, and H. Sahli, "Dcnn and dnn based multi-modal depression recognition," in *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 484–489.

[143] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews." in *Proceedings of INTERSPEECH*, 2018, pp. 1716–1720.

[144] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 43–50.

[145] M. Sharifa, R. Goecke, M. Wagner, J. Epps, M. Breakspear, G. Parker *et al.*, "From joyous to clinically depressed: Mood detection using spontaneous speech," in *Proceedings of the 25th International FLAIRS Conference*, 2012.

[146] N. Srimadhur and S. Lalitha, "An end-to-end model for detection and assessment of depression levels using speech," *Procedia Computer Science*, vol. 171, pp. 12–21, 2020.

[147] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 35–42.

[148] A. Vázquez-Romero and A. Gallardo-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy*, vol. 22, no. 6, p. 688, 2020.

[149] M. J. Patel, A. Khalaf, and H. J. Aizenstein, "Studying depression using imaging and machine learning methods," *NeuroImage: Clinical*, vol. 10, pp. 115–123, 2016.

[150] A. Wongkoblap, M. A. Vadillo, and V. Curcin, "Modeling depression symptoms from social network data through multiple instance learning," *AMIA Summits on Translational Science Proceedings*, vol. 2019, p. 44, 2019.

[151] A. Salekin, J. W. Eberle, J. J. Glenn, B. A. Teachman, and J. A. Stankovic, "A weakly supervised learning framework for detecting social anxiety and depression," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 2, pp. 1–26, 2018.

[152] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 3123–3128.

[153] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3–10.

[154] T. Giannakopoulos, "pyAudioAnalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[155] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[156] S. Mao, P. Ching, and T. Lee, "Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition." in *Proceedings of INTERSPEECH*, 2019, pp. 1686–1690.

[157] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[158] S. K. Pandey, H. S. Shekhawat, and S. Prasanna, "Attention gated tensor neural network architectures for speech emotion recognition," *Biomedical Signal Processing and Control*, vol. 71, p. 103173, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421007709

[159] C. De Boor and C. De Boor, *A practical guide to splines*. springer-verlag New York, 1978, vol. 27.

[160] J. Rottenberg, "Mood and emotion in major depression," *Current Directions in Psychological Science*, vol. 14, no. 3, pp. 167–170, 2005.

[161] J. A. Hall, J. A. Harrigan, and R. Rosenthal, "Nonverbal behavior in clinician—patient interaction," *Applied and preventive psychology*, vol. 4, no. 1, pp. 21–37, 1995.

[162] J. K. Darby and H. Hollien, "Vocal and speech patterns of depressive patients," *Folia Phoniatrica et Logopaedica*, vol. 29, no. 4, pp. 279–291, 1977.

[163] S. Harati, A. Crowell, H. Mayberg, and S. Nemati, "Depression severity classification from speech emotion," in *Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 5763–5766.

[164] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[165] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF

[166] S. Haq, P. J. Jackson, and J. D. Edge, "Audio-visual feature selection and reduction for emotion classification." in *Proceedings of AVSP*, 2008, pp. 185–190.

[167] J. James, L. Tian, and C. I. Watson, "An open source emotional speech corpus for human robot interaction applications." in *Proceedings of INTERSPEECH*, 2018, pp. 2768–2772.

[168] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[169] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[170] S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," in *Emotion measurement*. Elsevier, 2016, pp. 201–237.

[171] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, 2020.

[172] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[173] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[174] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2011.

[175] M. Senoussaoui, M. Sarria-Paja, J. F. Santos, and T. H. Falk, "Model fusion for multimodal depression classification and level detection," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 57–63.

List of Publications

Journal Publications

- Published:

  1. Sandeep Kumar Pandey, H.S.Shekhawat and S.R.M.Prasanna, "**Attention-Gated Tensor Neural Network Architectures for Speech Emotion Recognition**" (Biomedical Signal Processing and Control,Volume 71, Part A,Jan 2022)
  2. Sandeep Kumar Pandey, H.S.Shekhawat, S.R.M.Prasanna, Shalendar Bhasin, Ravi Jasuja "**A Deep Tensor Based Approach for Depression Recognition from speech utterances**", (PLOS ONE, August 2022).

- Manuscripts Communicated

  1. Sandeep Kumar Pandey, H.S.Shekhawat, S.R.M.Prasanna "**Multilingual Speech emotion recognition using deep neural embedding and metric learning**" (Biomedical Signal Processing and Control, *under review*)

Conference Publications

1. S. Pandey, S. Jelil, S. R. M. Prasanna and H. S. Shekhawat, " **Speaker Identification Using Tensor Decomposition of Acoustic Models**", TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South), 2018, pp. 1484-1488.

2. S. K. Pandey, H. S. Shekhawat and S. R. M. Prasanna, "**Deep Learning Techniques for Speech Emotion Recognition: A Review**", 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, Czech Republic, 2019, pp. 1-6.

3. Sandeep Kumar Pandey, H.S.Shekhawat and S.R.M.Prasanna, "**Emotion Recognition from Raw Speech using Wavenet.** ", IEEE TENCON 2019

4. Goel S, Pandey SK, Shekhawat HS. "**Analysis of Emotional Content in Indian Political Speeches**". In International Conference on Intelligent Human Computer Interaction 2020 Nov 24 (pp. 177-185). Springer, Cham.

5. Sandeep Kumar Pandey, Hanumant Singh Shekhawat, Shalendar Bhasin, Ravi Jasuja, and S R M Prasanna, "**Alzheimer's Dementia Recognition using Multimodal Fusion of Speech and Text Embeddings.** " International Conference on Intelligent Human Computer Interaction (IHCI) 2021, Kent, USA.

6. Ajit Kumar, Bong Jun Choi, Sandeep Kumar Pandey, Sanghyeon Park,SeongIk Choi, Hanumant Singh Shekhawat, Wesley De Neve, Mukesh Saini, SRM Prasanna, and Dhananjay Singh, "**Exploring Multimodal Features and Fusion for Time-Continuous Prediction of Emotional Valence and Arousal.**", International Conference on Intelligent Human Computer Interaction (IHCI) 2021, Kent, USA.

7. Sandeep Kumar Pandey, Mohit Nirgulkar, Hanumant Singh Shekhawat. "**A Longitudinal Study of the Emotional Content in Indian Political Speeches.**", International Conference on Intelligent Human Computer Interaction (IHCI) 2022, Tashkent, Uzbekistan.