DESIGN AND DEVELOPMENT OF FUSION FRAMEWORK USING PHONEMES AND MORPHEMES FOR SPOKEN WORD RECOGNITION

A Dissertation submitted to the Delhi Technological University for the Award of degree of

DOCTOR OF PHILOSOPHY
(INFORMATION TECHNOLOGY)

By
Sunakshi Mehra
2K19/PHDIT/02

under the Joint Supervision of

Dr. Ritu Agarwal
Assistant Professor
Department of IT

Dr. Virender Ranga
Associate Professor
Department of IT

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

February 2024

DECLARATION

I hereby declare that the dissertation titled "***Design and Development of Fusion Framework using Phonemes and Morphemes for Spoken Word Recognition***" that is being submitted by me for the **Doctor of Philosophy** degree is my own work and has not been submitted for the award of any degree or diploma to any other University or Institute. The work done in the thesis is original and has been done by me under the supervision of **Dr. Virender Ranga,** Associate Professor, Department of Information Technology, DTU, Delhi and **Dr. Ritu Agarwal**, Assistant Professor, Department of Information Technology, DTU, Delhi.

Place: Delhi Technological University, Delhi

Date:

This is to certify that the above statements made by the candidate are true.

Sunakshi Mehra
(2K19/PHDIT/02)

CERTIFICATE

This is to certify that the work contained in the thesis entitled "***Design and Development of Fusion Framework using Phonemes and Morphemes for Spoken Word Recognition***" submitted by Ms. Sunakshi Mehra (2K19/PHDIT/02) for the award of the degree of **Doctor of Philosophy** at Delhi Technological University, Delhi, India contains original research work conducted under our guidance and supervision.

We attest that Ms. Sunakshi Mehra has successfully met all the stipulated requirements and standards for the submission of the thesis. The research work presented in this thesis is original as per our best knowledge and has not been utilized as a foundation for the conferral of any other degree or similar recognition.

We, as the supervisor and joint supervisor, affirm the authenticity and originality of the research work conducted by Ms. Sunakshi Mehra and presented in this thesis.

Dr. Ritu Agarwal
Joint-Supervisor
(Assistant Professor)
Department of Information Technology
Delhi Technological University

Dr. Virender Ranga
Supervisor
(Associate Professor)
Department of Information Technology
Delhi Technological University

ACKNOWLEDGEMENT

This research journey would not have been possible with the help of many people. First and foremost, I would like to thank my supervisors **Dr. Virender Ranga** Associate Professor, Department of Information Technology, DTU, Delhi and **Dr. Ritu Agarwal**, Assistant Professor, Department of Information Technology, DTU, Delhi for their constant guidance and motivation. Their dedication to research has always encouraged me to work hard and strive for the best. I appreciate all their efforts, time, ideas, patience, and constant feedback on all the research papers. Furthermore, I would like to extend heartfelt gratitude to our Head of Department and DRC Chairperson, **Prof. Dinesh K. Vishwakarma**, Faculty members, my colleagues and lab members for their continue support. Their continuous support and invaluable suggestion have been instrumental in guiding me, whenever I needed assistance.

Finally, I would like to thank my parents, **Mr. Santosh Mehra** and **Mrs. Urmila Mehra**, my grandparents **Mr. Pyara Singh Mehra** and **Mrs. Channo Devi**, my guardians **Mr. Anil Kumar** and **Mrs. Anita Kumari**, my dearest husband **Mr. Amit Kumar**, for their unfailing love and support in all my pursuits. Without them, this journey wouldn't have been possible.

Place: Delhi Technological University, Delhi

Date:

Sunakshi Mehra
(2K19/PHDIT/02)

ABSTRACT

Spoken word recognition involves identifying words from spoken input. It specifically centers on recognizing and comprehending individual words within spoken language. Phonemes and morphemes play crucial roles in spoken word recognition. Phonemes are the smallest units of sound in a language, and they help differentiate between different words based on their pronunciation. Morphemes, on the other hand, are the smallest units of meaning in language, such as prefixes, suffixes, and root words. In spoken word recognition, phonemes help in distinguishing between words that sound similar but have different meanings. Morphemes provide additional context and meaning to words, aiding in understanding the overall message conveyed by the spoken language. By analyzing phonemes and morphemes, speech recognition systems can accurately identify and understand spoken words, enhancing their ability to convert spoken language into written text or commands. The study of spoken word recognition spans various fields like phonetics, linguistics, psychology, cognitive science, psycholinguistics, and computer science. Recent progress in deep learning and pre-trained models has transformed this field. These advancements allow for combining phonological and morphological parsing techniques, boosting the precision and effectiveness of recognizing words from spoken input. Uniting speech and understanding demands collaboration across multiple disciplines, reflecting the intricate and captivating nature of this research domain.

This thesis contributes to the advancement of spoken word recognition technology by providing a nuanced understanding of phonological and morphological features and offering a versatile fusion framework. The proposed system has potential applications in various fields, including speech processing, natural language understanding, and human-computer interaction. This thesis focuses on the design and development of a comprehensive fusion framework to improve spoken word recognition by integrating phonemes and morphemes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Definition
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BN	Batch Normalization
CMU	Carnegie Mellon University
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
DCT	Deep CO-Training
DL	Deep Learning
DOA	Direction of Arrival
DSR	Dysarthric Speech Recognition
DSUC	Dysarthric Speech Utterance Classification
DNN	Deep Neural Network
E2E	End-to-End
FDR	False Discovery Rate
FPR	False Positive Rate
FNR	False Negative Rate
GAN	Generative Adversarial Network
G2P	Grapheme-to-Phoneme
GFCC	Gammatone Frequency Cepstral Coefficients

GMM	Gaussian Mixture Model
GRU	Gated Recurrent Unit
GSCD	Google Speech Command Dataset
HCI	Human-Computer Interaction
HMM	Hidden Markov Model
IPA	International Phonetic Alphabet
LaBSE	Language Agnostics BERT Sentence Embedding
LNBNN	Local Naïve Bayes
LPCC	Linear Predictive Cepstral Coefficients
LRL	Low-Resource Language
LRW	Lip Reading in the Wild
LSTM	Long Short-Term Memory
MAML	Model Agnostic Meta-Learning
MCC	Matthews Correlation Coefficient
MFCC	Mel-frequency Cepstral Coefficients
ML	Multilingual
MOE	Mixture of Experts
MSA	Modern Standard Arabic
MSD	Multi-Space Probability Distribution
MSWC	Multilingual Spoken Words Corpus
NLP	Natural Language Processing
NMT	Neural Machine Translation
NN	Neural Network
NPV	Negative Predictive Value
PLP	Perceptual Linear Prediction

PPV	Positive Predictive Value
PNCC	Power Normalized Cepstral Coefficients
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristics
SA	Self-Attention
SCNN	Score-Based Convolutional Neural Network
SD	Speaker-Dependent
SGD	Stochastic Gradient Descent
SI	Speaker-Independent
SimMIM	A Simple Framework for Masked Image Modeling
SOTA	State-of-the-Art
STFT	Short-time Fourier Transform
SVM	Support Vector Machine
SWR	Spoken Word Recognition
TDNN	Time Delay Neural Network
TLS	Total Least Squares
TOM	Therapy Outcome Measure
t-SNE	t-Distributed Stochastic Neighbour Embedding
TTS	Text-to-Speech
USE	Universal Sentence Encoder
ViT	Vision Transformer
WER	Word Error Rate

LIST OF SYMBOLS

Symbols	Definition
b	Bias
B	Batch
β	Beta
b_f	Forget Gate Bias
c	Cell State
C	Category
fp	False Positives
fn	False Negatives
f	Forget Gate
h	Hidden State
i	Input Gate
l	Long-Term Component
o	Output Gate
p	Probability
pe	Expected agreement between raters
po	Observed agreement between raters
R	Rectified Linear Unit
X	Input
t	Time
$tanh$	Hyperbolic Tangent Function
tn	True Negatives
tp	True Positives

V	Variance
Var	Inference Variance
W	Weight Matrices
$\overline{Y}$	Output
γ	Gamma
Z	Activation Potential
$\rightarrow$	Forward Pass
$\leftarrow$	Backward Pass
α	Attention Weights
σ	Sigmoid
μ	Mean
Σ	Element-wise Summation
$*$	Element-wise Multiplication

CHAPTER 1

INTRODUCTION

This thesis contains an investigation of the speech which serves as an innate and effective mode of communication. We delve into the examination of spoken word recognition (SWR) analysis systems that employ fusion-based techniques to extract spoken words from speech cues in natural language. Our aim is to improve current speech recognition algorithms by incorporating fusion-based principles into the analysis of spoken words. This approach effectively addresses the inherent vagueness of natural language in a highly efficient and automated manner.

1.1 SPOKEN WORD RECOGNITION ANALYSIS- AN OVERVIEW

The inception of automatic speech recognition (ASR) can be traced back to 1952 when Peterson and Barney introduced this ground-breaking concept. This field represents an interdisciplinary convergence of computational linguistics, signal processing, and the artificial intelligence. Within this domain, the integration of intelligent devices equipped with speech recognition capabilities allows for the extraction of information from acoustic signals containing spoken language, facilitating the generation of user transcripts. These devices can also interpret various predefined expressions as commands, enabling control over their actions and functionalities. Research in the utilization of artificial intelligence techniques for speech detection and interpretation is strongly influenced by the human brain's innate speech processing mechanisms (Baber, 1991). The applications of Artificial Neural Networks (ANNs) in ASR gained traction during the 1980s due to their potential for parallel and distributed processing, as well as their ability to adapt to new acoustic patterns.

ASR, situated within the realm of artificial intelligence research, strives to develop systems capable of transforming the acoustic features of speech signals into digital sequences that can be represented as phonemes or written words (Daniel and James, 2009).

The primary objectives of ASR revolve around the advancement and implementation of efficient human-machine communication methods. Despite the higher computational demands inherent in verbal interactions compared to other interfaces like keyboards or touchscreens, ASR technologies offer advantage of quicker computational response times, especially for command-based operations. Furthermore, ASR proves to be an invaluable tool for individuals with physical-motor impairments (Damper, 1982), those whose occupations involve frequent manual tasks (Baber, 1991), and for the automation of both homes and businesses (Principi, Squartini, Bonfigli, Ferroni, and Piazza, 2015). The widespread adoption of ASR has encountered delays due to various challenges that can degrade the quality of the signal to be decoded by the system. Speech recognition is a complex task due to the myriad of factors affecting speech signals, such as background noise characteristics, speaker-dependent (SD) recognition, the phonetics and speech nuances of different languages, and even the semantic rules that influence the speech recognition process (Virtanen, Singh, and Raj, 2012). At its core, speech recognition is the process of transcribing spoken words into written text, where the fundamental elements of any spoken language are composed of its most basic symbols. These linguistic blocks, known as phonemes, serve as the foundational units of speech recognition.

The development of speech recognition systems necessitates a fundamental mechanism for recognizing phoneme units, as emphasized in Bhatt et al. (2020). In the context of speech recognition, the core components are the language model and the acoustic model. The performance of the acoustic model is intricately linked to the effectiveness of phoneme recognition, as noted in Nahar et al. (2016). To circumvent the challenges associated with handling large vocabularies, particularly given that words can be constructed by combining a language's phonemes, phoneme-based speech recognition is employed. However, it's important to acknowledge that this approach demands a substantially larger volume of training data compared to word-based models, primarily due to the lower number of

phonemes compared to words in each language. To mitigate the complexity of the system, neural network (NN), a prevalent tool in speech recognition systems, are leveraged. NN draw inspiration from the structural intricacies of the human nervous system and brain, as highlighted in Shrestha et al. (2019). These machine learning techniques have gained substantial attention owing to their innate capacity to extract latent features, enabling the manifestation of generalization across diverse recognition algorithms in various applications.

Alsulaiman et al. (2017) conducted a noteworthy study on the influence of Arabic phonemes on speaker recognition system performance. The study revealed varying identification rates for Arabic consonants, with certain combinations, such as a pharyngeal consonant followed by two nasal phonemes, achieving high recognition rates exceeding 80%, with the highest reaching 94%. Arabic vowels also displayed recognition rates surpassing 80%, while four additional consonants exhibited identification rates ranging from 70% to 80%. In the context of Arabic automated recognition systems, Alsharhan and Ramsay (2020) delved into data characteristics that can significantly impact system performance. Their experiments indicated a substantial reduction in the word error rate (WER) through the development of gender- and dialect-specific models. Herbig et al. (2011) demonstrated the creation of adaptive systems for unsupervised speaker tracking and speech recognition, while Malcangi and Grew (2017) introduced an innovative evolving connectionist method for adaptive audio-visual speech recognition. Finally, Koteswararao and Rao (2023) proposed a novel approach known as Multichannel KHMF, aimed at speech separation using enthalpy-based Direction of Arrival (DOA) and score-based convolutional neural network (SCNN) techniques. This approach has demonstrated efficacy in speech recognition, presenting new possibilities for enhancing recognition systems.

In the next chapter a comprehensive literature review is presented.

CHAPTER 2

LITERATURE REVIEW

2.1 IMPACT OF DEEP LEARNING ON SPOKEN WORD RECOGNITION

The recent advancements in deep learning (DL) techniques and the availability of extensive training datasets have led to significant improvements in speech recognition accuracy (Hinton et al. 2012). However, it has been observed that the robustness of deep neural network (DNN)-based models is highly contingent on the quality of the training data (Stern et al. 2008). Typically, the most substantial performance gains are achieved when the training data closely aligns with the test scenarios. Nevertheless, obtaining such a dataset may not always be practical. For instance, Google recently introduced Google Home (Li et al. 2017), a commercial product designed for far-field applications.

2.2 ANALYSIS ON ACOUSTIC FEATURES

Audio data is crucial in Human-Computer Interaction (HCI) across diverse domains. However, applying Speech Recognition to phone systems faces challenges requiring intelligent feature selection and robust DL algorithms. A comprehensive approach considers temporal and sequential emotional cues (Srivastava et al. 2014). Recurrent Neural Networks (RNNs) excel in speech-related tasks, but challenges like vanishing gradients hinder long-term correlation capture (Iqbal & Aftab, 2020). Dealing with exploding gradients involves truncation, while vanishing gradients demand holistic solutions. The term "speech

modulation spectrum" refers to the magnitude spectrum of the temporal envelope (Hermansky, 1997). Distorted temporal dynamics impact speech recognition. Contemporary ASR systems generate features at 100 Hz, incorporating speech dynamics using delta, double-delta features, or raw/smoothed speech temporal patterns for extensive context (Sadhu & Hermansky, 2023). ASR acoustic frontends commonly use time-frequency masking techniques (Narayanan & Wang, 2013). In a single-channel setup, a time-frequency mask estimates clean speech, forwarded to the ASR model in the complex spectral domain (Wang et al. 2020) or feature domain (Narayanan & Wang, 2013). Alternatively, frontend processing can occur directly in the time domain (Luo & Mesgarani, 2018). ASR frontend systems may inadvertently amplify speech due to misalignment with the training criterion, worsened by streaming constraints enforcing unidirectional and causal model architecture (Narayanan & Wang, 2013).

Growing interest in neural networks for speech enhancement involves DNNs, especially in regression-based speech enhancement (Xu et al. 2014). Parametric and non-parametric augmentation methods are utilized, with non-parametric lacking statistical richness and parametric potentially underutilizing the speech model (Chen, 2022). Augmentation techniques like speed perturbation (Liu et al. 2021) and tempo adjustment (Xiong et al. 2019) enhance data quality, with a crucial emphasis on developing a robust feature extractor for preserving ASR system recognition capabilities. Recent advancements introduce Gammatone Frequency Cepstral Coefficients (GFCC), outperforming Mel-frequency Cepstral Coefficients (MFCC), especially in noisy environments (Das & Bhattacharjee, 2014; Shuai et al. 2021). Power Normalized Cepstral Coefficients (PNCC) offer an intriguing alternative, demonstrating significantly improved accuracy in noisy conditions compared to MFCC (Kim & Stern, 2016). PNCC's efficacy is attributed to power-law non-linearity, asymmetric noise reduction, and temporal masking. In speech recognition feature extraction, PNCC utilizes a bank of Gammatone filters (Alzahra et al. 2017) to emulate non-linearly increasing bandwidths similar to the human auditory system. Convolutional Neural Networks (CNNs) further enhance feature extraction by emulating structural localization and minimizing translational variance within the feature space (Passricha & Aggarwal, 2019). For Uzbek and its dialects (Mukhamadiyev et al. 2022), a comprehensive approach involves

developing an End-to-End (E2E) DNN-HMM speech recognition model and a hybrid CTC-Attention network. This innovative method leverages the CTC objective function, reducing training time and improving speech recognition accuracy (Mukhamadiyev et al. 2022). To address variability in DNN-based ASR systems, feature transformation techniques like f-MLLR and VLTN have been introduced (Seide et al. 2011; Uebel & Woodland, 1999). Commonly employed are speaker-level f-MLLR transforms derived from Gaussian mixture model (GMM) -HMM-based ASR systems (Gales, 1998) (Seide et al. 2011). Acoustic features (Morshed & Ahsan, 2021) are vital for distinguishing speech classes, mitigating external noise, and managing speaker variability (Sadhu et al. 2019). Robustness against noise is a central concern in speech processing (Sankari et al. 2023). MFCC historically used for speech parameterization mimic human auditory characteristics, proving effective in tasks like speaker identification and speech recognition.

2.3 Spoken Word Recognition's Impact On Multilingual ASR's Potential

Numerous multilingual (ML) ASR models, including E2E and hybrid HMM/NN models, leverage data and parameters from all languages, enhancing robustness (Sercu et al. 2017). Training a unified model for all languages can benefit those with limited resources by transferring shared knowledge. Effective models often incorporate language information, demonstrated in research on ML representations (Ma et al. 2002) and E2E models (Watanabe et al. 2017). To adapt to data distribution variations, additional parameters can be introduced (Kannan et al. 2019). Adjusting language sampling ratios helps track shifts within an utterance (Jacobs and Bea, 1963). The expandability of ML models to newer languages is constrained by linguistic data dependence. This thesis explores challenges and opportunities in advancing speech recognition across diverse languages and scenarios. In low-resource language (LRL) ASR, adopting ML models, as emphasized by Ghoshal et al. (2013), proves promising. Recent projects like Babel (Tüske et al. 2013) and Spoken web search in MediaEval Benchmark (Metze et al. 2013) highlight ML model advantages. The Babel Optional Period 2 dataset is pivotal for studying speech recognition and keyword search technologies, evaluating ML representations (Plah et al. 2010). Developed independently by

IBM, Cambridge University, and RWTH Aachen, ML capabilities are a focus. While not exhaustive, strategies for expediting neural network training include distributed DNN training methods (Heigold et al. 2013) with significant communication costs. An alternative strategy by Sainath et al. (2014) uses a mixed hardware/software approach, and Seide et al. (2014) suggests 1-bit quantization of gradients to mitigate data communication costs. To expedite training, a strategy involves data sampling, as proposed by Byrd et al. (2011). Varying sample sizes in batch optimization methods, introduced in their framework, address large-scale machine learning challenges, contributing to accelerated training processes. This approach is relevant to the ongoing investigation in this thesis.

2.4 Examining The Role Of Phonology and Morphology In Shaping Spoken Word Recognition

Several techniques alleviate data requirements for LRLs in speech recognition. Benchmarks and probabilistic transcriptions (Glocker et al. 2023) are viable strategies, reducing the need for extensive training data. Architectural advancements enable zero-shot phoneme recognition for undiscovered languages, relying solely on phoneme inventories (Li et al. 2020). Allophone-to-phoneme mappings, crucial for handling language intricacies, are highlighted by Ladefoged (2014). In the 2010s, neural text-to-speech synthesis (TTS) introduces more natural and comprehensible speech output (Tan et al. 2021). Dealing with substantial data demands in LRLs poses challenges. A common approach is pre-training the acoustic model on a data-rich "source language" before fine-tuning on the limited data of the target LRL, as explored by Tu et al. (2019) and Wells and Richmond (2021). This study emphasizes employing neural networks for phonetics and phonology modeling, exploring unsupervised binary stochastic autoencoders (Shain and Elsner, 2019) and Generative Adversarial Networks (GANs) (Begus, 2020). While initially designed for vision and language categorization, these frameworks hold promise for language learning applications, including identifying phonemic tone contours in tonal languages (Li et al. 2019). This thesis focuses on advancing representation learning to address critical issues in speech recognition and historical phonology in LRL contexts.

Phonemes, the smallest units in speech, crucially distinguish word and sentence meanings, categorized as vowels and consonants. Consonants, with stronger airflow constraints, produce sounds with weaker amplitudes and noisier attributes (Rabiner and Juang, 1993; Deller et al. 1993). In contrast, vowels experience less airflow restriction. Arabic comprises 36 phonemes, including short and long vowels and consonants (Alghamdi, 2001). Understanding these distinct features is crucial for speech analysis and practical applications. Major languages, such as Japanese and certain American English variants, have varying vowel counts (at least five and 12, respectively). Bengali has 11 vowels, while Arabic employs a distinct pattern with three long and three short vowels, where vowel length is phonemically significant (Deller et al. 1993). Arabic's phonetic complexity includes diacritics serving as vowels. Consonant phonemes in Arabic may vary across dialects; for example, Egyptian Arabic replaces /ð/ and /Ø/ with /g/ for the letter /Z/ (Kirchhoff et al. 2002). Arabic dialects' diverse vowel and consonant phonemes pose intriguing challenges for linguistic analysis, showcasing the language's phonetic richness. Newman's (2002) study, comparing Arabic to other languages using the IPA framework and the UPSID database (317 languages, 58 phonetic features), focuses on the distinctive voiced pharyngeal fricative /è/. This rare phoneme is present in only eight languages, with five belonging to the Afro-Asiatic family. The prolonged variant, unique to Arabic and two other languages, holds significant importance in the study. Newman's investigation focused on Arabic's pharyngeal and uvular phonemes, highlighting the rare occurrence of sounds like the voiced pharyngeal fricative, showcasing Arabic's uniqueness. Brazilian Portuguese, stemming from Portuguese, boasts a complex phonological system with 36 phonemes, including 26 consonants and 10 vowels. Brazilian Portuguese features phonetic representations for 21 oral diphthongs and 5 nasal vowels, enhancing its phonological richness within the Romance language family (Silva & Yehia, 2011). The incorporation of vowel digraphs and nasal vowels makes Brazilian Portuguese linguistically distinctive, with back vowels conveying crucial information through lower frequencies, and front vowels exhibiting distinct characteristics in the higher frequency range. Nasal vowels represent a subset of vowel phonemes. Nasal vowels, produced by opening the oral cavity and lowering the soft palate for airflow through both nasal and oral cavities (Bisol, 2005), contribute to unique acoustic output. This behavior defines back and front vowels, along with the distinctive qualities of nasal vowels within the

phonological system. In natural speech, pauses rarely occur within individual words, with prosodic events, marked by tonic fluctuations, influencing both acoustic and semantic attributes based on preceding and following words (O'Shaughnessy, 2008). Nasal vowels, with pronounced weightiness in lower frequencies, exhibit longer durations compared to oral vowels. Enunciating nasal vowels reveals an initial band transitioning into the nasal sound, impacting initial resonance frequencies (Theodor et al. 1994). In the realm of vowel groups, diphthongs play a significant role as monosyllabic speech sounds, capable of transitioning either upwards or downwards in sound. Rising diphthongs emerge when an additional vowel co-occurs in the same syllable with /i/ and /u/, starting as a semivowel and progressing into a fully pronounced vowel. In summary, nasal vowels are characterized by weighty, low-frequency attributes, while diphthongs are essential components of vowel groups, showcasing sound variations that can ascend or descend based on specific vowel combinations. The performance of an ASR system is significantly influenced by the physical environment, including the recording area, devices for sound capture, and communication channels. The type of microphone used, particularly its frequency response within the human speech range (100 Hz to 8000 Hz), plays a pivotal role in capturing precise characteristics for successful transcription. Selecting an appropriate microphone with optimal frequency response is crucial for peak ASR system performance. The International Phonetic Alphabet (IPA) serves as a standardized system for accurately representing phonetic elements across diverse languages (Association, 1999).

2.5 Enhancing Spoken Word Recognition For Asian Languages With Transformers

ASR technology is a vital tool for transcribing and understanding spoken language, with applications in voice user interfaces, dictation, interactive voice response systems, and language learning platforms. It significantly enhances social interaction and accessibility, particularly for individuals with disabilities, promising to simplify daily life. In information technology, ASR is widely used in voice-driven applications, automatic language translation, and various voice-operated systems (AbuZeina et al. 2011). Arabic speech recognition research faces challenges due to limited data, extensive lexical diversity, numerous spoken

dialects, and the complexity of Arabic's morphology. Developing acoustic models for dialectal Arabic poses a significant obstacle, requiring training with the specific dialect used (Boumehdi and Yousfi, 2022). Creating a comprehensive pronunciation dictionary for dialectal Arabic words is exceptionally challenging due to the absence of standardized spellings in spoken dialects. Diacritization for Arabic dialects is more intricate than for Modern Standard Arabic (MSA), requiring a dialectal Arabic morphological analyzer for generating various diacritization forms. The absence of a robust language model for dialectal Arabic complicates context-based diacritization. Challenges arise due to a larger set of vowels in dialectal Arabic, particularly in aligning them with audio input (Abdou and Moussa, 2019). European language literature explores various subword modeling techniques, including morphological analyzers and n-grams for subword dictionary acquisition. Subword ASRs utilizing morphs and rule-based algorithms have been studied by Hirsimäki et al. (2006) and Byrne et al. (2000). Morphological analyzers feature in works by Erdogan et al. (2005), Laureys et al. (2002), and Hacioglu et al. (2003). Hacioglu et al. (2003) and Arısoy et al. (2007) developed subword ASRs using data-driven algorithms. Kirchhoff et al. (2006) amalgamated morphological and lexical information for factored and joint lexical morphological language models, including direct use of morpheme-based n-grams. Creutz and Lagus's (2005) Morfessor 1.0 is a renowned approach for morpheme segmentation and morphology induction. Additionally, TF-IDF, an earlier technique, is recognized as a potent lexical feature (Metze et al. 2013). In English speech recognition research, widely used corpora include Switchboard (300 hours), LibriSpeech (960 hours), TedLium-3 (450 hours), Common Voice (1400 hours), and SPGISpeech (5000 hours) (O'Neill et al. 2021). Except for Switchboard, these corpora are freely available for academic and non-profit use. Recent years have seen a substantial reduction in WER in English speech recognition, reaching 1.4% on the test-clean LibriSpeech dataset (Panayotov et al. 2015; Amodei et al. 2016; Zhang et al. 2020). However, there is a significant challenge in low-resource Indian languages, lacking a comparable large speech corpus for ASR research. Addressing this, Microsoft Billa (2018) released a dataset tailored for low-resource Indian languages, providing 50 hours of speech transcriptions in Tamil, Telugu, and Gujarati, totaling 150 hours of data. Each language has 40 hours of training data and 5 hours of test data, supporting advancements in ASR technology for these underrepresented languages.

Efforts to address resource limitations in languages include data augmentation techniques demonstrated by Liu et al. in 2019. Creating a substantial speech corpus in Tamil is crucial for transfer learning and pre-trained models. Van Huy et al. (2014) introduced the Multi-Space Probability Distribution Hidden Markov Model (MSD-HMM) for Vietnamese speakers, simulating phonemes and incorporating tone information. Using Perceptual Linear Prediction (PLP) and MFCC acoustic features with four unique streams, the Vietnamese MSD-HMM improved accuracy by 2.49% compared to the best baseline system and 0.54% compared to the best system without MSD-HMM. Markov models, highlighted by He and Ferguson in 2022, enhance the efficiency and reliability of speech systems. Tone is crucial in tonal languages like Mandarin and Thai, as highlighted by Hu et al. in 2014. Nguyen (2016) introduces the use of Scale-Invariant Feature Transform (SIFT) for voice categorization, combining it with Local Naïve Bayes (LNBNN) for speech categorization. This approach demonstrates significantly higher accuracy in classifying speech signal properties. In neural network-based continuous expression identification with a vast vocabulary, Gehring et al. (2017) propose a modular synthesis approach.

Deep bottleneck technologies significantly enhance DBN/HMM hybrid systems' productivity, as noted by Sainath et al. (2012). Their research shows up to a 21.5% relative improvement over the MFCC baseline. In Vietnamese speech recognition, Nga et al. (2021) introduce an E2E approach with Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber (1997)) and TDNN models, demonstrating superior accuracy through deep learning. Adaptation to specific language databases while maintaining simplicity for future developments remains a key challenge. To enhance ML modeling, modifications to neural network models have been proposed. Pratap et al. (2020) demonstrated a multi-decoder ML model, while Kannan et al. (2019) used adapter layers in encoder-decoder models to address varying data availability. Zhu et al. (2020) and Pratap et al. (2020) parameterized attention heads per language in Transformer-based encoders. Mixture of Experts (MOE) models, introduced by Jacobs et al. (1991), are common in speech recognition and machine translation (Papi et al. 2021; Jain et al. 2019). Lu et al. (2020) applied MOE to a bilingual code-switching system, utilizing pre-trained monolingual encoders for effective E2E ML ASR. MOE models consist of multiple expert models, each specialized for specific data

subsets, integrated using a gating function for effective overall performance. Training ML ASR models on large-scale data encounters linguistic differences. An experimental approach introduces two challenges: context-free phoneme sequence prediction addresses orthographic disparities, and domain adversarial classification enhances adaptability to language-specific phenomena (Ganin et al. 2016). Prior work by Krishna et al. (2018) combined grapheme and phoneme objectives, but the scale of language adversarial training in ASR, as highlighted by Yi et al. (2018), is unprecedented. Speaker-related challenges, influenced by emotions, gender, or age, pose complexities in multi-modal analysis and fusion (Wu et al. 2022). Variation in accents within the same language, exemplified in Arabic, adds another layer of complexity (He and Ferguson, 2022).

2.6 Advances In Spoken Word Recognition On Speech Impairment Datasets

This section provides an overview of the literature on Dysarthric Speech Recognition (DSR) in datasets characterized by strong cognitive qualities. It delves into conventional speech enhancement and recognition techniques, weighting their respective advantages and disadvantages. Classical ASR models often employ HMMs, Support Vector Machines (SVMs), ANNs. However, DL techniques have excelled in achieving the SOTA. Individuals with dysarthria, a motor speech disorder resulting from brain damage, often struggle with speech recognition technology designed for typical speakers. This is due to limited availability and inadequate training for those with dysarthria, relying mainly on statistical acoustic models (Dhanalakshmi et al. 2018). Dysarthria, linked to neurological conditions like Parkinson's disease, cerebral palsy, strokes, or head injuries, hinders speech production with symptoms such as weakness and coordination deficits. Assistive technology aids dysarthric individuals in communication and device control, addressing associated physical disabilities (Xie et al. 2022; Whitehill & Ciocca, 2000; Enderby, 2013; Hawley et al. 2012, 2007).

Advances in DNN topologies, demanding substantial training data, further exacerbate the issue. Dysarthria is prevalent among older individuals, especially those with neurocognitive

conditions like Alzheimer's disease, impacting their speech patterns (Ye et al. 2021). DSR proposes a two-step model adaptation process—initially creating a speaker-independent (SI) dysarthria model and then refining a SD dysarthria model (Takashima et al. 2020). This approach adapts shared knowledge from SI non-dysarthria models, demonstrating superior performance compared to traditional methods. Dysarthria, arising from conditions affecting speech motor control, involves muscle weakness and coordination issues, leading to inconsistent vowel and inadequate consonant articulation. Preferred over generic systems (Enderby, 2013), specialized speech recognition models for dysarthric speakers account for these distinctive characteristics (Trinh et al. 2022). In 1995, Jayaram and Abdelhamied pioneered a hybrid HMM-ANN dysarthric-specific ASR system using a dataset of dysarthric individuals with mild speech intelligibility. While typical speech recognition excels, dysarthric speakers face limited access and training, relying on statistical acoustic models (Dhanalakshmi et al. 2018). DNNs demand substantial training data, posing challenges in dysarthria data collection. The UA-Speech database enhances ASR training, outperforming traditional models across dysarthria severity (Kim et al. 2008). The HMM-based dysarthric-specific recognizer achieves a reported 69.1% recognition rate (Selouani et al. 2009), lacking detailed WERs and speech intelligibility assessment. In their 2010 study, Dede and Sazlı leveraged ANNs to achieve an impressive 78.25% recognition rate for dysarthric ASR. Despite the challenge of a mere 20% speech intelligibility, this breakthrough underscores ANNs' prowess in enhancing ASR, particularly in the context of dysarthric speech. ASR, prevalent in smartphone personal assistants, relies on deep learning, demanding ample training data. Traditional systems, trained on standard speech, often struggle with dysarthric speech due to challenges in acquiring sufficient data from individuals with articulation difficulties (Xiong et al. 2018; Shahamiri, 2021). Athetoid symptoms in dysarthric individuals further complicate recognition. Vachhani et al. 2018) synthesized dysarthric speech from healthy speech, classified using a random forest classifier. Dysarthric classes trained DNN-HMM-based ASR systems, and a triple-speed perturbation approach enhanced data, expanding the dataset (Xiong et al. 2020). Factored TDNN-F, incorporating CNNs, achieved an average WER of 30.76% on the UA single-word speech corpus. Dash & Solanki (2020) proposed a modified DNN for speech enhancement, improving signal quality but maintaining computational costs. Rahiman et al. (2021) utilized DL for channel estimation

in ambient noise conditions with enhanced reliability. Haridas et al. (2018) employed the delta-AMS method for noise removal, adjusting the delta parameter for varying noise levels. Woszczyk et al. (2020) developed an adversarial neural network improving voice recognition across severity levels for 11 individuals. The fusion-based speech analysis methodologies devised in this thesis serve the purpose of assessing spoken language across audio, video, and textual data. Each of the challenges, spanning multimodal speech analysis ad speech recognition, hinges upon the application of fusion-based algorithms. These algorithms leverage audio, image, and text features to discern and identify spoken language within speech data. Solutions to these challenges are achieved through the implementation of supervised, unsupervised, and hybrid classification algorithms, incorporating fusion techniques that amalgamate phoneme and morpheme-based features.

2.7 THE INSPIRATION FOR TACKLING THE ISSUES EXPLORED IN THE THESIS

Nonetheless, the multitude of accents within our globalized society and the presence of ambient noise present considerable challenges in automatically identifying speech from real-world samples. The necessity for robust, streamlined, and genuine interaction has surged due to the expanding adaptive landscape of voice interfaces in intelligent devices (Ravuri & Stolcke, 2015; Tur & Mori, 2011). Within this context, the significance of natural language processing (NLP) is paramount for discerning text and speech category labels (Lin et al. 2021). Yet, the copious volumes of ML speech data accessible online render voice search and unstructured dictation inadequate for encompassing the linguistic spectrum (Cabrera et al. 2021). In conjunction with NLP, acoustic characteristics extracted from raw audio, such as MFCC, are routinely harnessed for speech recognition Davis & Mermelstein (1980). ASR entails converting spoken language into readable text through computational means. This endeavour encounters difficulties due to the fact that diverse speakers exhibit varying styles, accents, and vocal attributes, making it intricate to derive precise text transcripts from audio recordings (Besacier et al. 2014). Recent times have witnessed substantial advancements in supervised classification techniques for transcribing spoken content, whereas comparatively less emphasis has been placed on exploring unsupervised methodologies. In recent times,

voice-activated devices have garnered widespread popularity and have seamlessly integrated into our daily routines. Prominent illustrations include Google Home, Amazon's Alexa, and Apple's Siri. These intelligent systems harness ASR to transform spoken language into text. However, transcribing spoken discourse into coherent written form presents a range of challenges. These encompass the frequent presence of inaccuracies within the generated transcriptions, which can significantly influence the quality of interactions with virtual assistants (Ogawa & Hori, 2017). Moreover, although encoding words into character strings and converting them into vectors via embedding techniques proves effective for refined text, this approach falls short when dealing with transcriptions extracted from audio recordings (Fang et al. 2020). As a remedy, a dependable strategy for robust speech recognition emerges through the direct extraction of phonemes from the original audio recordings. This becomes especially vital in situations involving speech that is accented or affected by noise. A widely embraced method for phoneme extraction from raw audio involves the use of PocketSphinx (Huggins-Daines et al. 2006).

Numerous research studies have underscored the substantial influence of accents on ASR accuracy. For instance, Feng et al. (2021) conducted experiments that illuminated significant variations in system accuracy for native speakers in different geographical locations, primarily attributed to accents. These experiments have demonstrated that incorporating embeddings as additional inputs can effectively capture accent- and speaker-related information, leading to a significant improvement in the accuracy of accented ASR systems (Tong et al. 2022). Additionally, specialized datasets like SUBAK.KO (Kibria et al. 2022) have been introduced to capture the diverse regional accented pronunciations, further emphasizing the importance of modeling acoustic variabilities introduced by accents in the domain of speech recognition (Xie et al. 2022). An intriguing finding, as highlighted in (Dokuz and Tüfekci, 2022) suggests that combining gender and accent features contributes to the enhancement of speech recognition performance. One notable technique for improving accented speech recognition involves the learning of mappings between accented and canonical phones. These mappings can be generated through various means, including subject-matter expertise (Richards and Schmidt, 2013), extensive exposure to accented and unaccented speech (Goronzy et al. 2004; Loots and Niesler, 2011), or the development of

mappings under specific hand-crafted constraints (Humphries et al. 1996). These strategies play a crucial role in addressing the challenges posed by accents in speech recognition, a key focus of investigation within this thesis.

2.8 RESEARCH GAPS

The current landscape of research in the field of speech processing reveals notable gaps and opportunities for advancement. Specifically, there is a noticeable deficit in comprehensive studies that integrate multiple speech frameworks. Researchers have not given sufficient attention to the classification of audio-phonemes and text-phonemes concurrently, representing an untapped potential for understanding the intricate relationship between spoken and written language. Furthermore, the realm of feature extraction and classification is underexplored when considering both raw audio and speech transcriptions simultaneously. This holistic approach could lead to more robust and nuanced models, capable of capturing the nuances present in both auditory and linguistic representations. In addition, the intersection of phoneme and morpheme classification remains an underdeveloped area of inquiry. Combining these linguistic units in classification tasks has the potential to provide a deeper understanding of the interplay between sound and structure in language. Addressing these gaps in research would not only contribute significantly to the theoretical foundations of speech processing but also hold practical implications for the development of more accurate and versatile language models.

By fostering interdisciplinary collaboration and encouraging exploration in these overlooked areas, researchers can pave the way for ground-breaking advancements in the understanding and application of speech and language processing technologies.

2.9 RESEARCH OBJECTIVES

In order to systematically address the identified research gaps in the field of speech processing, we propose a comprehensive set of research objectives designed to explore and

contribute to the unexplored intersections within the domain. The overarching aim of our research is to enhance the understanding and capabilities of speech processing models. Our specific objectives include:

1. Conduction of spectrogram-based phonological studies for spoken word recognition (Chapter-3).
2. Design and development of fusion framework for phoneme-based spoken word recognition from raw audio (Chapter-4).
3. Design and development of fusion framework for phoneme- and morpheme-based spoken word recognition from speech transcriptions (Chapter-5).
4. Design and development of classification framework for phonological and morphological features using pre-trained networks for spoken word recognition (Chapter-6).
5. Design and development of fusion framework for spoken word recognition from raw audio and speech transcriptions (Chapter-7).

2.10 THE CHALLENGES EXAMINED WITHIN THIS THESIS AND THE CORRESPONDING RESOLUTIONS

As evident from the preceding section, there is a compelling need for advanced fusion-based algorithms within the context of Speech Analysis. These algorithms should possess the capability to accurately assess both phonological and morphological information while effectively addressing the inherent ambiguities present in natural language. Following the introduction, the thesis is organized into seven chapters. Chapter 2 entails an extensive literature review concerning the SOTA developments in the Speech field. Chapter 3 through 7 delve into innovative methodologies for conducting speech analysis, specifically tailored to tackle the imprecision and uncertainty inherent in natural language.

Briefly outlined below are the fusion-based solutions proposed within this thesis to address the five identified problems in spoken word analysis.

- **Problem 1: A new fusion technique for highlighting of spectrogram-based phonological studies for spoken word classification is to be developed.**

In Chapter 3, we propose a novel approach to investigates the relatively unexplored roles of spectrograms and phonology in enhancing the precision of SWR. To achieve this goal, a novel dual-pronged approach is employed, utilizing the Speech2Text transformer to separately handle text transcript extraction and spectrogram generation. The experimentation phase is conducted using the Google Speech Command Dataset (GSCD), which encompasses both 10-word and 35-word categories. In order to create two-dimensional audio representation, mel spectrogram images (Sakashita & Yono, (2018) are processed and resized to dimensions of 256 x 256 pixels. These images are subsequently categorized using both the ImageNet and the tiny Swin transformer version 2. Additionally, a grapheme-to-phoneme (G2P) model is incorporated to convert Speech2Text transcripts into phonemes. Through a method called phoneme slicing, essential phonological features like fricatives, nasals, liquids, glides, plosives, approximants, taps/flaps, trills, and vowels are extracted, with careful consideration of factors such as manner and place of articulation. The study aims to decode spoken words accurately, assessing spectrogram and phonological impact through ablation analysis. It introduces a late fusion strategy, combining phone and image embeddings, achieving impressive accuracy. This surpasses existing methods, setting a new benchmark by integrating linguistic insights. The approach harmonizes text and mel spectrograms, enhancing ASR precision and highlighting the vital role of phonological analysis in speech interpretation.

- **Problem 2: A new fusion framework that highlights spoken word classification and facilitates phoneme-based spoken word recognition from raw-audio.**

In Chapter 4, we present a supervised approach for the recognition of accented speech, particularly when dealing with limited resources. One notable gap in previous research is the underutilization of phonology in understanding spoken text. Our proposal involves the early fusion of phone embeddings to effectively identify accented speech, even when working with a small sample dataset. We begin by extracting phonemes from .wav recordings using

PocketSphinx. These phonemes are then transformed into vectors using FastText's character n-gram-based subword embeddings. To ensure uniformity, we concatenate and pad the vectors. The early fusion of phone embeddings yields an impressive accuracy, when applied to the task of accented speech recognition, specifically in the context of the L2-ARCTIC accented speech corpus. This accuracy surpasses that of existing techniques. Our work aims to demonstrate the significant role that audio phonemes can play in accented speech recognition, even in scenarios with a limited number of training samples.

- **Problem 3: A new fusion framework that emphasizes both spoken word classification and the recognition of spoken words based on phoneme and morpheme information extracted from speech transcriptions.**

In Chapter 5, we present an unsupervised approach designed to enhance the accuracy of speech transcriptions that are initially highly imperfect. This improvement is achieved through decision-level fusion, involving stemming and a two-way phoneme pruning process. Our transcripts are obtained from videos by first extracting audio using the Ffmpeg framework and subsequently converting the audio into text transcripts using the Google API. The benchmark dataset used for evaluation is the Lip Reading in the Wild (LRW) dataset, comprising 500 word categories, each with 50 videos in mp4 format. Each video consists of 29 frames, each lasting 1.16 seconds, with the target word positioned in the middle of the video. To enhance the baseline accuracy, we applied various techniques including stemming, phoneme extraction, filtering, and pruning. The stemming algorithm is applied to the text transcript, resulting in a notable accuracy for word recognition. For phoneme conversion, we utilized the Carnegie Mellon University (CMU) pronouncing dictionary, which provides a phonetic mapping of English words to their pronunciation. Our two-way phoneme pruning process phonemes containing vowels and fricatives. After implementing stemming and two-way phoneme pruning, we incorporated decision-level fusion techniques, ultimately achieving a substantial improvement in word recognition rates, reaching an accuracy. This approach significantly enhances the accuracy of transcribing highly imperfect speech.

- **Problem 4: A novel fusion framework that highlights both spoken word classification and recognition, encompassing classification frameworks for phonological and morphological features using pre-trained networks for spoken word recognition.**

In Chapter 6, we present a supervised technique for Spoken Word Recognition (SWR), especially when dealing with limited input data. Notably, previous research has shown limited exploration of the potential of morphemes and phonemes in understanding spoken text. To address this, we introduce a late fusion approach involving phone embeddings and bigrams embeddings to identify spoken words from a small sample collection. Our audio recordings are stored in .OPUS format, and we extract text transcripts from the raw audio using the English Large xlsr-Wav2Vec2-53 pre-trained classifier. We recover phonemes from the text transcript using the CMU pronouncing dictionary and convert them into vectors through ML, language-agnostic sentence embeddings. To ensure uniformity, we concatenate and pad the vectors. Additionally, we extract bigrams from the text transcript and vectorize them using the same ML language-agnostic sentence embeddings. Both the phoneme embeddings and morpheme embeddings are input into a 5-layered dense batch normalization model.

The results indicate that the late fusion of phone embeddings and bigrams achieves a good accuracy in Arabic, Vietnamese, and Tamil for the 10 spoken word categories in the ML spoken words corpus. These accuracies surpass those of existing techniques. Our work underscores the significant role that linguistics, specifically phonemes and morphemes, can play in SWR, even when dealing with limited and imbalanced training samples. We also propose that text-transcription features extracted from pre-trained models outperform existing audio-based feature modeling.

We also developed a supervised strategy tailored for SWR in a ML dataset under resource constraints. Addressing the paucity of research exploring the application of morphology and phonology in comprehending spoken text is a primary focus, setting it apart from SOTA approaches. The ML spoken words corpus stores audio files in .opus format. To obtain text

transcripts from the original audio, we deploy the pre-trained Arabic Large xlsr-Wav2Vec2-53 transformer. Our experiment unfolds in two stages, involving two distinct forms of text transcripts: “buckwalter transliteration” and “Arabic script”. In the initial stage, we convert the buckwalter transliteration form into phonemes by leveraging the CMU pronouncing dictionary with the support of an Arabic-based grapheme-2-phoneme model. These phonemes are then translated into vectors using character-n-grams-based subword embeddings provided by FastText. Moving to the second stage, Arabic scripts undergo stemming, followed by conversion into unigrams. FastText word embeddings facilitate the transition from unigrams to vectors. In both scenarios, we concatenate and pad the vectors to ensure uniformity.

These vectors, collected in both stages, feed into a three-layered dense model with batch normalization to generate probabilistic scores. The results are calculated by averaging the outcomes of both stages, yielding satisfactory results that outperform SOTA approaches.

- **Problem 5: An innovative fusion framework that accentuates both spoken word classification and recognition, encompassing the design and development of spoken word recognition from raw audio and speech transcriptions.**

In Chapter 7, our study tackles the challenge of effectively integrating multimodal data from imperfect text transcripts and raw audio within a deep framework for automatic speech recognition. Our approach emphasizes late fusion of audio and text modalities. We introduce a SA-deep BiLSTM model to independently process audio and text data. For training each feature type, we employ the SA-deep BiLSTM model, which consists of five BiLSTM layers and incorporates a self-attention module between the third and fourth layers. We consider linguistic data, such as word stems extracted from text transcripts, and acoustic features like Mel MFCC and Mel-spectrograms. The GloVe word embedding is utilized to vectorize linguistic data.

By combining the posterior class probabilities obtained from SA-deep BiLSTM models trained on individual modalities, we achieve an impressive accuracy on the 10-word

categories of the Google speech command dataset. Rigorous testing using this dataset and ablation analysis demonstrate the superiority of our proposed method, primarily due to the consistently high classification accuracies it achieves compared to SOTA approaches.

In the upcoming chapter, we have explored objective 1, delving into the impact of spectrogram and phonology on the recognition of spoken words.

CHAPTER 3

CONDUCTION OF SPECTROGRAM-BASED PHONOLOGICAL STUDIES FOR SPOKEN WORD RECOGNITION

Simulating the task of identifying words in human speech is challenging due to the complex nature of the human mind's cognitive processes. In this chapter, we propose an advance speech recognition by effectively integrating multimodal data, including text transcripts and mel spectrograms extracted from raw audio. The study investigates the relatively unexplored roles of spectrograms and phonology in accurately recognizing spoken words. To accomplish this, we adopt a dual approach by utilizing the Speech2Text transformer to separate the acquisition of text transcripts and spectrogram extraction. We conduct experiments on the GSCD, which includes both 10-word and 35-word categories. We process mel spectrogram images, resizing them to 256 × 256 pixels to create two-dimensional audio representations. These images are then categorized using the ImageNet and the tiny Swin transformer version 2.[1]

[1]
The contents of this chapter are submitted/accepted/under review in:
"Evaluating the significance of suprasegmental features in speech command recognition through spectrogram and phonological fusion analysis" – Soft Computing. (IF: 3.1).
&
"Improving speech command recognition through decision-level fusion of deep filtered speech cues" - Signal, Image and Video Processing SIViP (2023), https://doi.org/10.1007/s11760-023-02845-z. (IF: 2.3).
&
"A Deep Learning Approach to Dysarthric Utterance Classification with BiLSTM-GRU, Speech Cue Filtering, and Log Mel-Spectrograms" –The Journal of Supercomputing (2024), https://doi.org/10.1007/s11227-024-06015-x. (IF: 3.3).

Additionally, we employ a G2P model to convert Speech2Text transcripts into phonemes. Our primary objective is to accurately decode the linguistic information embedded in spoken words. To evaluate the impact of spectrograms and phonological characteristics on categorization, we conduct an ablation analysis. We propose a late fusion strategy that combines phone embeddings and image embeddings. Our strategy outperforms other methods, setting a new benchmark by merging linguistic insights with abundant resources. By bridging the gap between text transcripts and mel spectrograms, our approach provides a robust solution for achieving more accurate and reliable ASR performance. This work contributes to the field of ASR by exploring the synergies of multimodal data in a technical and comprehensive manner. The remaining sections of this objective are structured as follows: In Section 3.1, we delve into the pre-processing of data transcripts. Section 3.2 is dedicated to audio pre-processing and the extraction of features. In Section 3.3, we explore phonology, including the study of suprasegmental phonemes. Moving on to Section 3.4, we discuss phoneme embeddings. Our proposed methodology is presented in Section 3.5. Experimental details and discussions are covered in Section 3.6. Section 3.7 provides an in-depth ablation analysis of spectrogram and phonology. Finally, Section 3.8 summarizes the key findings of this chapter.

3.1 SPEECH2TEXT TRANSCRIPTS PRE-PROCESSING

As outlined by Alsharhan and Ramsay (2019), a phoneme, the smallest unit of speech in linguistics, is characterized by distinct sounds or groups of sounds that carry variations in meaning and pronunciation. These pronunciation differences can be influenced by adjacent letters, affecting their representation in written language. In our methodology, we leverage text transcripts, as depicted in Figure 3.1, to extract these phonemes. To accomplish this, we employ a pre-trained transformer model, “Speech2Text”, to extract text transcripts from audio samples. Subsequently, we utilize a G2P model to convert the remaining transcripts into phonemes. The CMU Pronunciation Dictionary, a comprehensive resource with over 125,000 words and their associated phonetic transcription, serves as the foundation for representing the possible speech sounds of words. This representation encompasses critical aspects such as stress, articulation, and intonation. To enhance clarity and precision, we

engage in text normalization, generating a set of phonemes from the aligned transcript. Our focus is on capturing the intricate sound patterns of syllables and the stress patterns within words or phrases. Notably, we opt for typical phonetic representations for each word, omitting numerical stress values. For example, the word “backward” is phonetically transcribed as “ “B”, “AE1”, “K”, “W”, “ER0”, “D” ”. In this representation, stress is indicated by numerical values, where 0 denotes no stress, 1 signifies primary stress, and 2 indicates secondary stress. To retain the essential phonetic information, we eliminate these numerical stress markers. As illustrated in Figure 3.1, the transcript undergoes segmentation into individual words, with each word represented by its corresponding phonemes.

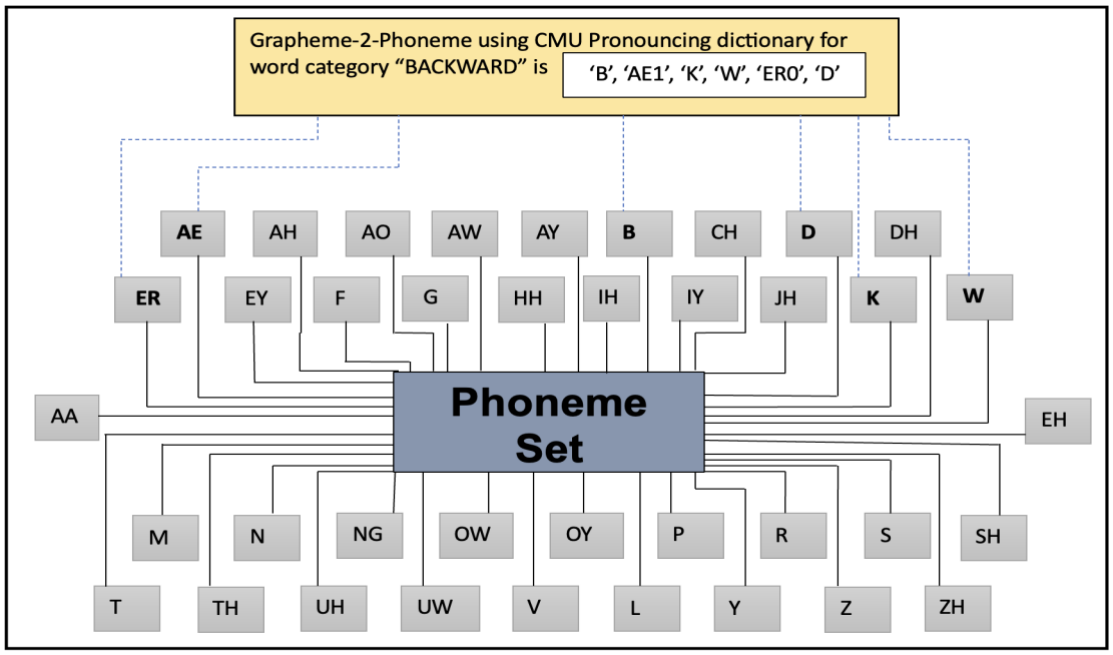


Figure 3.1. Using a G2P model, the text transcription of the word "backward" is analyzed to determine the appropriate phonemes that represent the sounds in the word

However, it’s important to acknowledge that achieving an exact match in transcripts can be challenging due to the inherent variations in pronunciation resulting from different speakers (Hazen, 2006). Our approach relies on the CMU Pronouncing Dictionary, tailored for American English phonetic representation, and adept at capturing stress patterns.

Nevertheless, we remain attentive to the ongoing challenge posed by speaker-related variations. The dictionary aligns with the IPA, ensuring a consistent representation of spoken language sounds, encompassing vowels, plosives, nasals, glides, and fricatives. Plosives, often referred to as stop or oral consonants, exert temporary blockage of the vocal tract, resulting in the cessation of airflow. This category encompasses both voiced and voiceless consonants. Among the voiced plosives, we find the letters “b”, “d”, and “g”, whereas voiceless plosives consist of the letters “p”, “t”, and “k”. Distinguished by their pronounced amplitude, notable fricatives encompass the letters “f”, “s”, “v”, and “z”.

Table 3.1 The International Phonetic Alphabet's letters (Association, 1999)

	Bilabial	Labio-dental	Dental	Alveolar		Post-alveolar	Retro-flex	Palatal	Velar	Uvular	Pharyngeal	Epiglottal	Glottal
Plosives	p, b				t, d		ú, ã	c, Í	k, g	q, å		Ü	P
Fricatives	f, b	f, v	T, D	s, z		S, Z	ù, ü	ç, J	x, G	X, K	è, Q	Ý, Ë	h, H
Affricates						Ù, tZ							
Taps				R			ó						
Approximants		V		ô			õ	j	î				
Vibrants	à			r						ö			
Lateral fricatives				ð , ì									
Lateral flaps				Õ									
Nasals	m	M		n			ï	ñ	N	ð			
Lateral approximants				l			í	ń	Ï				

The paramount aim of the IPA is to offer a comprehensive representation of sounds prevalent in all spoken languages. It accomplishes this by assigning unique identifiers and categories to each linguistic unit on the phonetic chart, a fundamental introduction by the IPA as shown in Table 3.1. Consonants primarily originate within the vocal tract, particularly the oral tract, which encompasses the mouth and pharynx. Based on their specific points of articulation, consonants are further classified into dorsal, labial, coronal, and radical categories. In contrast, vowels contribute to elevate pitch (Bastanfard et al. 2009) and amplitude, rendering them generally distinguishable and easily detectable. The distinction between nasal and oral speech is contingent upon the presence or absence of airflow blockage in the oral tract. Fricatives, such as “f”, “s”, “v”, and “z”, are produced by pressing the lower lip against the upper teeth to create these sounds. Trills and taps, created by active and passive articulators, respectively, introduce two additional acoustic elements (Schwartz and Makhoul, 1975). In

summary, the process of phoneme filtering, coupled with the utilization of the CMU Pronouncing Dictionary ensures a consistent representation of speech sounds. Consonants characterized by their high intensity amplitude, often referred to as glottal stops, are classified as fricatives.

3.2 Audio Pre-Processing And Feature Extraction

The initial phase of signal pre-processing is a crucial step in ensuring accurate speech recognition. It servers to enhance key components of incoming sound signals, thereby improving their acoustic attributers proceeding to feature extraction. In this context, the application of a median filter proves valuable for the removal of undesirable noise elements from input audio signals. The median filter operates by replacing each data point within the signal with the median value of its neighbouring data points, utilizing a predefined window size specified as the kernel_size. The kernel size, represented by filter_size, is set to 5, effectively removing noise and enhancing the quality of the speech signal through median filtering (Mehra et al. 2023). Utilizing a median filter for speech signal enhancement offers several advantages, including noise reduction (Rabiner and Schafer, 2007), preservation of speech features, robustness against outliers, simplicity, efficiency, non-linear noise reduction, edge preservation, adaptability, and compatibility with .wav format (Gonzalez and Woods, 2008). Subsequent to background noise removal via the median filter, the improved speech signals are organized into separate categories and stored in distinct folders. After this pre-processing step, Mel spectrograms are generated from the speech audio samples. The Mel spectrograms are extracted using the Librosa package, involving the application of the short-time Fourier transform (STFT) to the spoken signal. This process includes the conversion of signal amplitudes into decibels and the mapping of frequencies onto the Mel scale. Consequently, it yields the Mel spectrogram, offering a visual representation of the signal's spectral content. These enhanced Mel spectrograms are then input into the Swin transformer, a powerful deep-learning model detailed further below. The Swin transformer employs three key techniques to optimize its performance: Residual-post-norm with Cosine Attention: This method combines a residual-post-norm architecture with cosine attention, contributing to enhanced training stability. The utilization of the residual-post-norm approach promotes smoother gradient flow during training, while the cosine attention mechanism bolsters the model's capacity to capture meaningful relationships across different input segments. Log-spaced continuous position bias: The Swin transformer employs a log-

spaced continuous position bias, facilitating the effective capture of long-range dependencies. This technique empowers the model to attend to distant elements within the input, ultimately augmenting its grasp of the broader context within the Mel spectrograms. Self-supervised pretraining A Simple Framework for Masked Image Modeling (SimMIM): To reduce reliance on large labeled image datasets, the Swin transformer employs the self-supervised pretraining approach known as SimMIM. The Swin transformer leverages SimMIM to acquire meaningful representations from unlabeled data by predicting image patch orders. It categorizes Mel spectrograms with knowledge drawn from 14 million annotated ImageNet images, thereby enabling efficient handling of image-to-image tasks. Its proficiency in processing Mel spectrograms underscores its potential in the area of speech signal processing and speech-related applications. The flowchart depicted in Figure 3.2 provides a visual representation of the proposed approach.

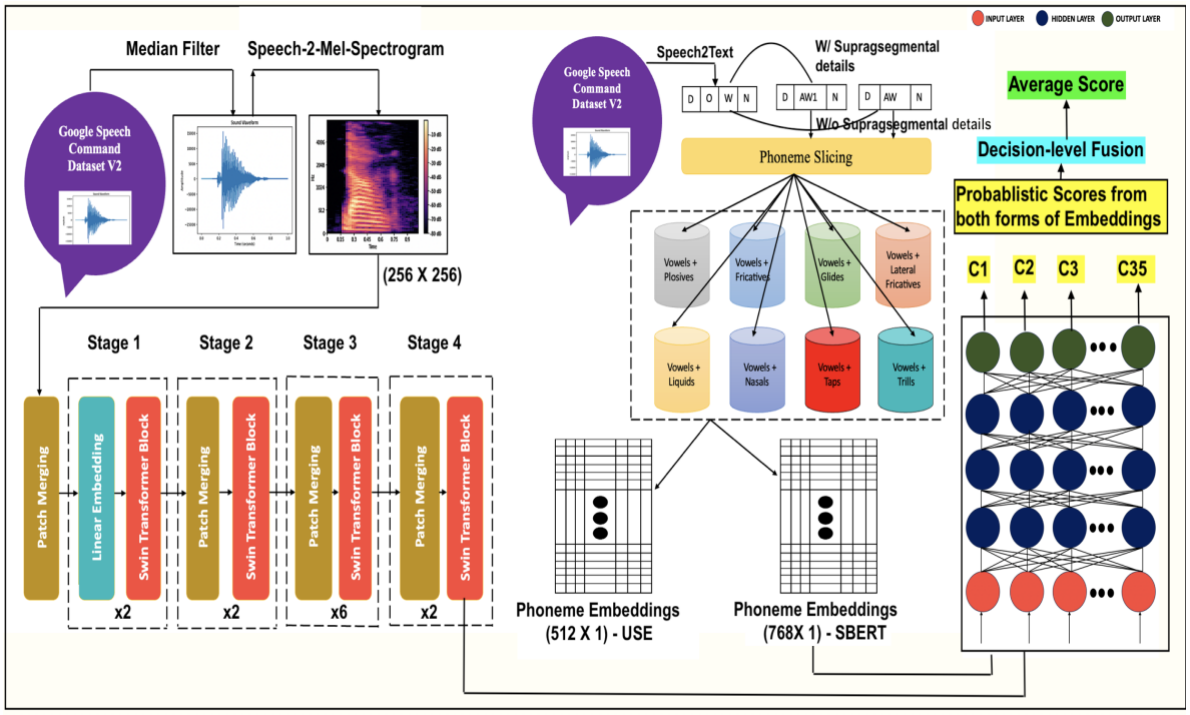


Figure 3.2. Workflow of proposed approach

3.3 With And Without Suprasegmental Phoneme

Suprasegmental phonemes, as discussed by Yenkimaleki and Heuven (2021), are representations that encompass the prosodic and rhythmic elements of speech, including

intonation, stress, and rhythm. These phonemes extend beyond individual phonemes and encapsulate the characteristics of entire speech units or phrases. Incorporating suprasegmental phonemes into speech-processing tasks can notably enhance the model's capacity to comprehend and generate speech that sounds natural. Their utilization in speech-processing tasks enables models to capture the subtleties of spoken language more effectively, leading to improved performance across various speech-related applications such as speech recognition, speech synthesis, and spoken language understanding. This integration of prosodic information empowers the model to generate speech output that is both natural and expressive. The taxonomy of suprasegmental stress levels, as depicted in Figure 3.3, allow for the classification of stress in speech. When suprasegmental phonemes are omitted, models may solely rely on segmental phoneme information, potentially resulting in speech generation that sounds less natural. This could lead to difficulties in capturing the rhythmic and expressive aspects of speech, culminating in a more robotic or monotonic speech output (Sönmez and Varol, 2021). In summary, the inclusion of suprasegmental phoneme information significantly enhances the model's ability to produce expressive and authentic speech, rendering it more akin to human speech. This supplementary prosodic information proves valuable in various speech-related contexts, particularly in conveying emotions, emphasis, and rhythm. Our analysis encompassed both scenarios with and without suprasegmental phonemes, evaluating their contributions to accurately identifying spoken utterances.

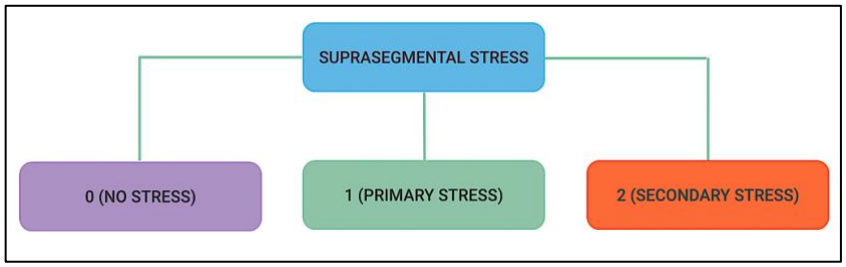


Figure 3.3. Levels of stress in spoken utterance

3.4 PHONEME EMBEDDINGS

Utilizing the Universal Sentence Encoder (USE) (Cer et al. 2018), we conducted the generation of phoneme embeddings. This widely acclaimed pre-trained model, developed by Google, excels at producing consistent fixed-length vector representations for sentences and texts. In our context, we applied this model to phonemes, both with and without

suprasegmental additions, tailored to the spoken utterance category. This yielded embeddings of size (512 × 1). The USE, based on the Transformer architecture, is highly versatile for a spectrum of NLP applications. Its strength lies in its extensive training on diverse textual data, allowing it to effectively capture the semantic meaning of phrases. On another front, we harnessed Sentence-BERT (SBERT) embeddings, an alternative class of pre-trained models engineered to create high-quality fixed-length vector representations (768 × 1) for sentences and texts. In our study, SBERT played a pivotal role in extracting embeddings for both suprasegmental and non-suprasegmental phonemes. SBERT, an extension of BERT (Bidirectional Encoder Representations from Transformers), a renowned language model for word embeddings, operates by fine-tuning BERT for sentence-level tasks through the inclusion of sentence pairs during training. Unlike traditional BERT, where masked language modeling is the primary objective, SBERT employs Siamese and triplet networks alongside contrastive loss functions to master sentence representations. The resultant embeddings obtained from the USE and SBERT encapsulate critical phonological attributes, such as fricatives, vowels, and plosives. Their expressiveness is further enriched through processing via a three-layered feed-forward network, facilitating the capture of intricate speech patterns. The depth of this network leverages hierarchical phonetic features to construct a discriminative feature space. Meticulous training, incorporating activation functions and regularization, ensures the effective harnessing of phonological information. The output of this network evolves into a potent representation for downstream speech-related tasks, including phoneme recognition, speech synthesis, and accent classification. The amalgamation of phonemes and SBERT reinforces both phonological and semantic representations, culminating in enhanced speech processing performance.

3.5 3-Layered Neural Network Architecture

The integrated audio and image embeddings are input into a three-layered dense model, as depicted in Figure 3.4. These features are processed using the three-layered model, comprising a flattened layer and three dense layers with 512, 256, and 64 units, each employing the ReLu activation function. The learning rate, regulated by the Adam optimizer (Duchi et al. 2011), is controlled in a stochastic gradient descent (SGD) fashion. This optimizer stands out for its consistent performance across a variety of tasks. The loss function employed is known as sparse categorical cross-entropy.

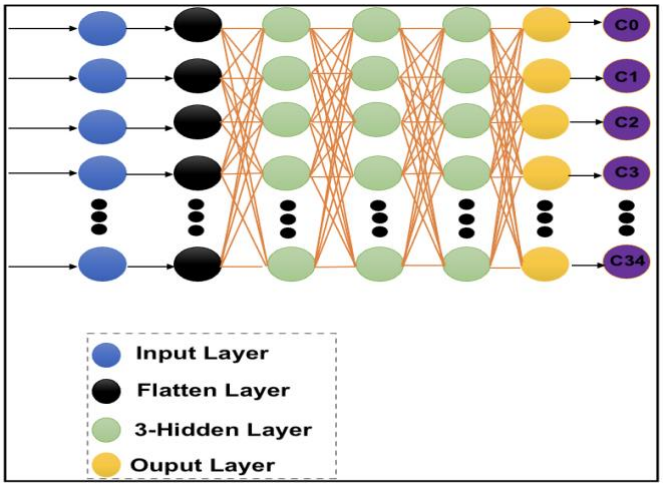


Figure 3.4. 3-layered dense model for classifying the fused image and linguistic embeddings

3.6 EXPERIMENTAL DETAILS AND DISCUSSIONS

GSCD version 2 is a widely used collection, featuring 105,937 one-second audio recordings of 35 spoken commands, including "yes," "no," "stop," and more. Each clip maintains a consistent sampling rate of 16 KHz and follows a single-channel (mono) format. This dataset is crucial for training and evaluating machine learning models in speech recognition, keyword spotting, and wake-word detection tasks. The 35-word category occupies 13.87 GB, while the 10-word category variant is 1.25 GB. Researchers and developers leverage its diverse commands for advancing speech-related applications. Curated by Google, this dataset stands out for its diverse audio recordings, encompassing various speakers, accents, and age groups. The 35 spoken commands are recorded in different environments, ensuring robustness to environmental variations. It is divided into training (80-90%), validation, and test sets for comprehensive model evaluation. Each audio clip is labeled with its corresponding command and organized into folders, making it a valuable resource for voice-controlled apps, wake-word detection, and speech recognition in real-world scenarios.

Table 3.2 Summary of dataset

CATEGORY	TRAIN	VALIDATION	TEST
RIGHT	3778	363	396

GO	3880	372	402
NO	3941	406	405
LEFT	3801	352	412
STOP	3872	350	411
UP	3723	350	425
DOWN	3917	377	406
YES	4044	397	419
ON	3845	363	396
OFF	3745	373	402
EIGHT	3787	346	408
CAT	2031	180	194
TREE	1759	159	193
BACKWARD	1664	153	165
LEARN	1575	127	161
BED	2014	213	207
HAPPY	2054	219	203
DOG	2128	197	220
WOW	2123	193	206
FOLLOW	1579	132	172
NINE	3934	356	408
THREE	3727	356	405
SHEILA	2022	204	212
ONE	3890	351	399
BIRD	2064	182	185
ZERO	4052	384	418
SEVEN	3998	387	406
VISUAL	1592	139	165
MARVIN	2100	195	195
TWO	3880	345	423
HOUSE	2113	195	191
SIX	3860	378	394
FIVE	4052	367	445
FORWARD	1557	146	155
FOUR	3728	373	401

The summary of dataset is presented in Table 3.2. This dataset plays a pivotal role in the training and assessment of machine learning models for diverse speech-related tasks, including speech recognition, keyword spotting, and wake-up word detection. These tasks are crucial for ensuring optimal performance in real-world scenarios. However, it is noteworthy that the dataset comes with its share of challenges. These challenges include background noise, variations in pronunciation, and commands that sound similar. Addressing these challenges is paramount for the development of robust and highly accurate models. The dataset finds widespread application in voice-controlled applications, wake-

word detection in smart devices, and speech recognition for voice-activated virtual assistants.

3.6.1 HYPERPARAMETERS

In our investigation, we conducted extensive testing using Python software version 3.10.0 on a macOS Big Sur platform, equipped with an M1 chip. To ensure the accessibility of our code for future studies, we made it available online. Given the computationally intensive nature of our experiments, which involved the use of audio samples and transformers, we opted for Google Colab Pro with GPU support, alongside the Librosa library to facilitate the experimental computations. It's important to note that Colab Pro offers a maximum of 32 GB of RAM, which is instrumental in handling our tasks. Our initial steps involved generating Mel spectrogram images with dimensions of (256×256) to serve as inputs to the Swin transformer. Subsequently, we applied a three-layered neural network model to the pre-trained probabilistic scores, incorporating translation, rescaling, and resizing operations to prepare the images for the transformer. In our experimental framework, we utilized pre-trained probabilistic scores derived from the Swin transformer as input to a feed-forward neural network. The training regimen spanned 100 epochs to effectively capture underlying data patterns. To introduce nonlinearity and augment the model's ability to grasp intricate relationships within the data, we harnessed the Rectified Linear Unit (ReLU) activation function. The choice of ReLU has exhibited its effectiveness in a myriad of classification tasks. For efficient regulation of the learning rate during stochastic gradient descent, we adopted the Adam optimizer, a well-established choice proposed by Duchi et al. in 2011. Renowned for its consistent performance across diverse classification tasks, the Adam optimizer computes current gradients by considering an average of previous gradients, thereby promoting stable and efficient updates during training. The use of the Adam optimizer is instrumental in optimizing our neural network's convergence and enhancing its generalization capabilities, even on unseen data. Within the scope of our research, we harnessed the Speech2Text transformer to extract textual transcripts from audio samples. Initially, these transcripts are represented as graphemes, which are written symbols representing English words. However, our primary objective is to perform the conversion of these graphemes into their corresponding phonemes, reflecting the accurate pronunciation of words. To accomplish this crucial G2P conversion, we relied on the CMU pronouncing dictionary, a valuable linguistic resource introduced by Hazen in 2006. This dictionary offers

a comprehensive mapping of English words to their respective phonetic representations. Leveraging this resource, we achieved a highly precise transformation of text transcripts into their corresponding phonemes. This process of G2P conversion holds significant relevance in a variety of language-related tasks, encompassing areas such as speech recognition, natural language integration, and linguistic analysis. Having access to phonetic representations of words enables us to gain a deeper understanding of pronunciation patterns, stress markers, and various phonological features present in spoken text. The IPA defines an array of factors, including places of articulation, stress markers, sound articulation, production methods, and features related to voiced and voiceless sounds, fricatives, glottal sounds, and stop/plosive sounds. Our approach not only enhances the precision of transcript representations but also paves the way for further exploration in fields like keyword spotting, word correction, and the reduction of WERs. Additionally, this transformation empowers us to construct a hierarchical lexical framework that encompasses both morphological and phonological elements, ultimately facilitating a more comprehensive understanding of linguistic structures within the audio data.

3.7 ABLATION STUDY

In this section, we conducted an ablation study to investigate the impact of audio-based and linguistic features, as well as their combination, on a dense model. Our proposed technique emphasizes the influence of both audio-based and linguistic factors on speech command recognition. Our study reveals a clear advantage in classification scores through the fusion of audio-based and linguistic information. Additionally, incorporating stress and intonation details into linguistic features positively impacts classification scores. This section covers a comparison with other SOTA approaches and discusses the individual impacts of audio-based and linguistic-based features on spoken command classification. A comprehensive comparison of our methodology using the Swin-T embedded 3-layered model with SOTA techniques is presented in Tables 3.3 and 3.4.

Table 3.3 Performance evaluation against the SOTA for the Google Speech Command Dataset's 10-word category

Comparison for 10-word categories	ACC (%)
MFCC + CNN (Haque et al. 2020)	93.28%

GFCC + CNN (Abdelmaksoud et al. 2021)	93.09%
MFCC + LSTM-RNN (Wazir et al. 2019)	95.44%
MFCC + LSTM-RNN (Zia and Zahid, 2019)	95.14%
MelSpec + LSTM (Lezhenin et al. 2019)	95.07%
DenseNet + BiLSTM (Zeng and Xiao, 2018)	94.88%
RNN neural attention (de Andrade et al. 2018)	94.11%
EdgeCRNN (Wei et al. 2021)	98.20%
Semi Supervised audio tagging (Cances and Pellegrini, 2021)	95.58%
Attention based s2s model (Higy and Bell, 2018)	97.50%
TripletLoss-res15 (Vygon and Mikhaylovskiy, 2021)	98.38%
BC-ResNet-8 (Kim et al. 2021)	98.70%
KWT-3 (Berg et al. 2021)	98.49%
MatchboxNet-3x2x64 (Majumdar and Ginsburg, 2020)	97.63%
ConvMixer (Ng et al. 2022)	98.21%
Embedding + Head (Lin et al. 2020)	97.70%
Wav2KWS (Seo et al. 2021)	98.52%
Proposed approach (Spectrogram + Phonemes-SBERT)	99.87%

Table 3.4 Performance evaluation against the SOTA for the Google Speech Command Dataset's 35-word category

Comparison for 35-word categories	ACC (%)
TripletLoss-res15 (Vygon and Mikhaylovskiy, 2021)	97.00%
KWT-3 (Berg et al. 2021)	97.69%
RNN neural attention (de Andrade et al. 2018)	93.90%
M2D (Niizumi et al. 2023)	98.50%
Eat-S (Gazneli et al. 2022)	98.15%
AST (Gong et al. 2021)	98.11%
HTS-AT (Chen et al. 2022)	98.00%
ImportantAUG (Trinh et al. 2022)	95.00%
KW-MLP (Morshed and Ahsan, 2021)	97.56%
Proposed approach (Spectrogram + Phonemes-SBERT)	98.53%

The results, which attain the highest test accuracy of 99.87%, undeniably demonstrate the superiority of our technique in a well-resourced data environment. Our approach outperforms a CNN employing MFCCs as input by a margin of 6.59% (Haque et al. 2020) for 10-spoken command categories. When compared to alternative methods, our technique achieves a higher accuracy of 4.43% compared to Mel spectrogram with LSTM (Wazir et al. 2019). Our method surpasses EdgeCRNN (Wei et al. 2021), which is a feature-enhanced approach based on depth-wise separable convolution and residual structure, by 1.67%. Moreover, despite the presumed effectiveness of GFCCs in detecting emotions, our technique outperforms the CNN-GFCCs approach (Abdelmaksoud et al. 2021) by 6.78%.

Additionally, our approach excels in comparison to DenseNet-BiLSTM (Zeng and Xiao, 2018), a model recommended for keyword spotting, achieving a remarkable accuracy of 94.88%. Furthermore, our method outperforms several commonly used designs. For speech command recognition, the accuracy scores using different models are as follows: TripletLoss res 15 (Vygon and Mikhaylovskiy, 2021) achieves 98.38%, BC ResNet 8 (Kim et al. 2021) attains 98.70% accuracy, Self-attention keyword spotting transformer KWT 3 (Berg et al. 2021) scores 98.49%, RNN neural attention (de Andrade et al. 2018) reaches 94.11%, MatchboxNet 3×2×64 (Majumdar and Ginsburg, 2020) records 97.63%, ConvMixer (Ng et al. 2022) obtains 98.21%, keyword spotting with Embedding + Head (Lin et al. 2020) achieves 97.70% and Wav2KNWS (Seo et al. 2021) attains 98.52% accuracy for 10-word categories. This concise summary is presented in Table 3.3. Our model significantly outperforms other transformer-based models. For speech recognition with 35-word categories, the accuracy scores using different models are as follows: TripletLoss res15 (Vygon and Mikhaylovskiy, 2021) achieves 97.00%, KWT 3 (Berg et al. 2021) reaches 97.69% accuracy, RNN neural attention (de Andrade et al. 2018) scores 93.90%, M2D (Niizumi et al. 2023) obtains 98.50%, Eat S (Gazneli et al. 2022) reaches 98.15%, AST: Audio spectrogram transformer (Gong et al. 2021) records 98.11%, HTS AT: a hierarchical token semantic audio transformer (Chen et al. 2022) achieves 98.00%, ImportantAUG (Trinh et al. 2022) scores 95.00%, and KW MLP (Morshed and Ahsan, 2021) attains 97.56% accuracy for 35-word categories. In summary, our proposed approach outperforms SOTA methods for categorizing both 10-word and 35-word categories on the Google Speech Command dataset, achieving excellent accuracy rates of 99.85% and 98.01%, respectively. These outstanding results can be attributed to the dataset's attributes, which include SD samples with a variety of accents, and the ideal data sample quantities per category. This concise summary is presented in Table 3.4. Visualizations using t-Distributed Stochastic Neighbour Embedding (t-SNE), as presented in Figures 3.5 and 3.6, underscore the superiority of spectrograms over phonological analysis extracted from text transcripts, for both 10-word and 35-word categories. The t-SNE visualization illustrates the influence of the Swin Transformer on mel spectrogram images extracted from the Google Speech Command dataset. Spectrogram-based clusters are notably distinct and well-separated, indicative of the exceptional classification achieved with the Swin-T transformer model. The t-SNE visualization showcases the model's capacity to effectively capture and represent intricate patterns and features within the spectrogram data. The clear separation of clusters

further emphasizes the model's ability to discern subtle differences among speech samples, leading to highly accurate and reliable classifications. This study highlights the potential for significant advancements in speech recognition by leveraging the power of spectrogram-based representations and sophisticated transformer models.

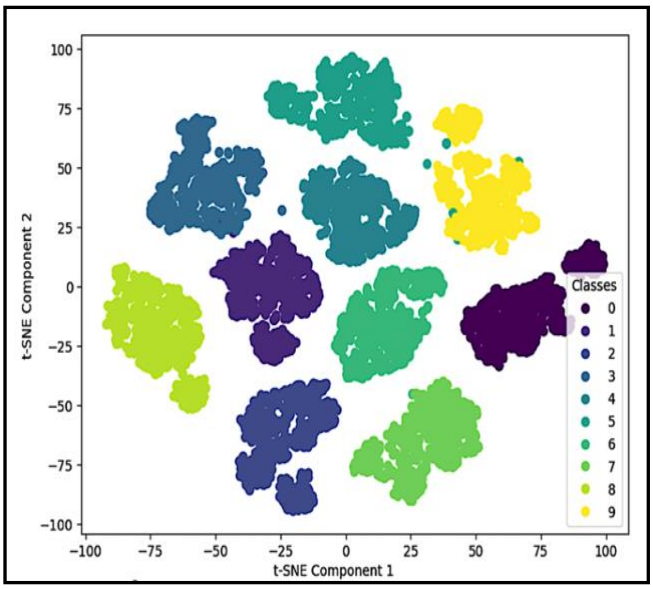


Figure 3.5. 3-layered dense model for classifying 10-word categories using Swin-T transformer embeddings

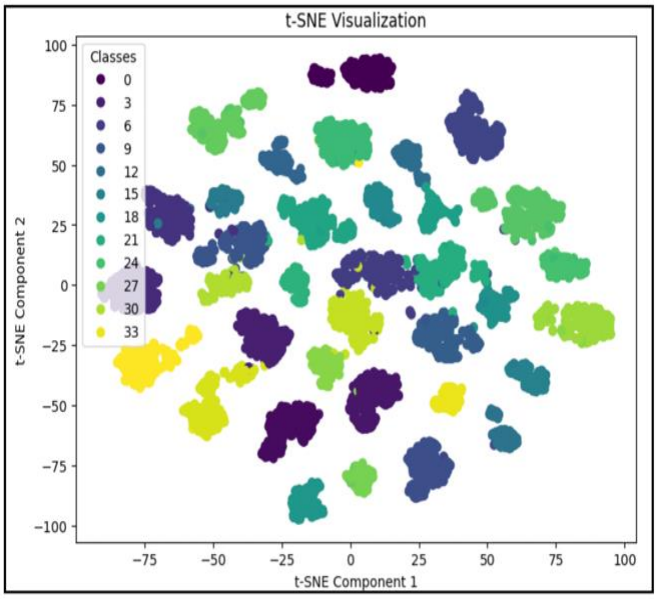


Figure 3.6. 3-layered dense model for classifying 35-word categories using Swin-T transformer embeddings

Figures 3.7 to 3.10 illustrate the training and testing accuracy results concerning phonological attributes, both with and without stress markers, utilizing Sentence BERT and USE embeddings in conjunction with a 3-layered neural network. The analysis unequivocally demonstrates that the inclusion of stress markers consistently leads to enhanced accuracy for both Sentence BERT and USE embeddings. Stress markers emerge as pivotal elements in capturing the nuanced aspects of speech, culminating in more precise phonological representations and significantly improved classification performance.

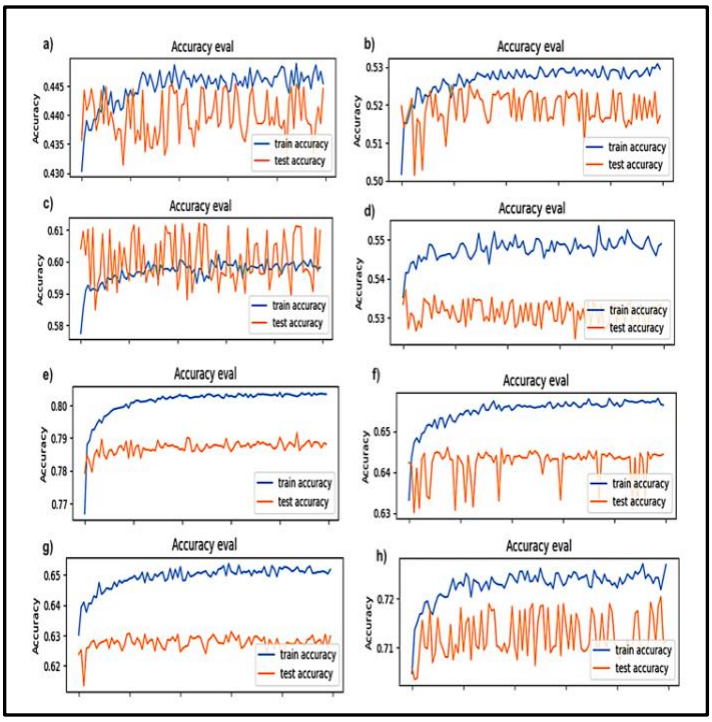


Figure 3.7. 3-layered dense model for classifying 10-word categories using phonology with stress markers where a) lateral fricatives, b) nasals, c) fricatives, d) glides, e) plosives, f) traps, g) liquids and h) trills using 512 USE embeddings

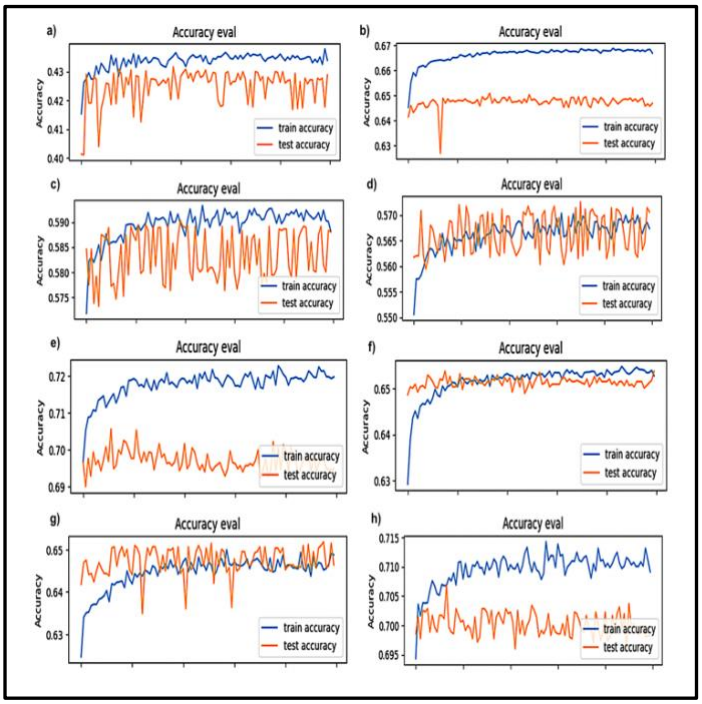


Figure 3.8. 3-layered dense model for classifying 10-word categories using phonology without stress markers (0,1, and 2) where a) lateral fricatives, b) nasals, c) fricatives, d) glides, e) plosives, f) traps, g) liquids and h) trills using 512 USE embeddings

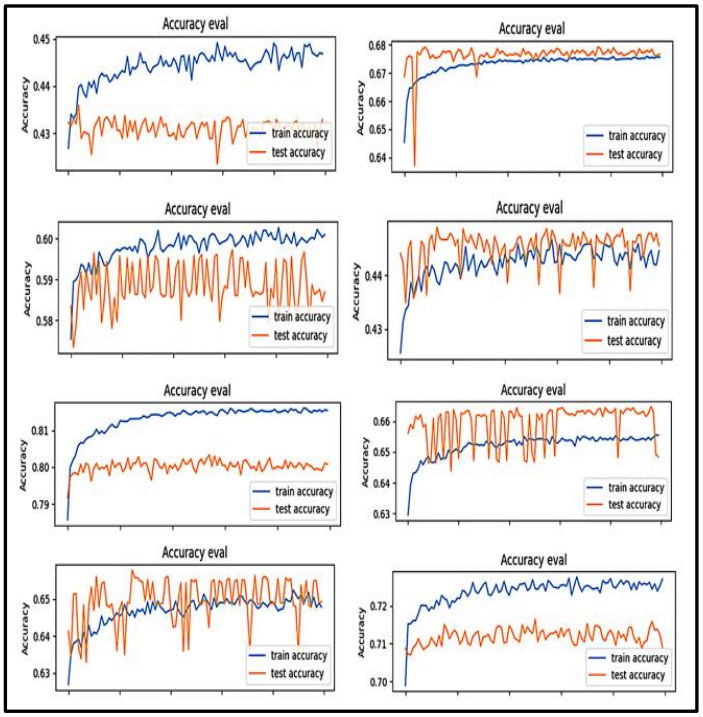


Figure 3.9. 3-layered dense model for classifying 10-word categories using phonology with stress markers where a) lateral fricatives, b) nasals, c) fricatives, d) glides, e) plosives, f) traps, g) liquids and h) trills using 768 Sentence-BERT embeddings

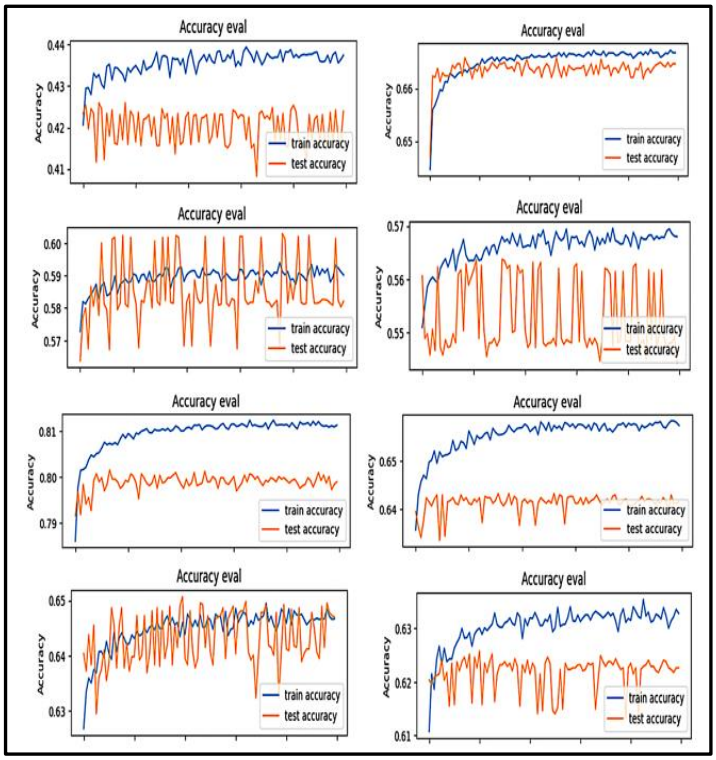


Figure 3.10. 3-layered dense model for classifying 10-word categories using phonology without stress markers where a) lateral fricatives, b) nasals, c) fricatives, d) glides, e) plosives, f) traps, g) liquids and h) trills using 768 Sentence-BERT embeddings

Within the framework of our proposed in-depth analysis and decision-level fusion approach, which incorporates our dense model, we scrutinized the significance of individual audio/visual components, pairwise combinations of male and female speakers, and the deep model itself. The key findings of the ablation analysis, summarized in Tables 3.5 - 3.11, are as follows, illustrating the ultimate performance for both audio-based and linguistics-based classification within each class. Notably, the combined classification of audio and textual features outperforms the suggested dense framework.

- In the proposed dense framework, image-based classification exhibits superior performance compared to text-based classification.
- Our work employs dual-branch audio and visual modalities, and these outperform individual combinations of other widely-used audio/visual components.
- A significant discovery is that distinct speakers, achieve high accuracy rates, notably 99.85% for image-based classification and 90.64% for spoken word classification using phonemes (without stress markers) on SBERT embeddings. When stress markers are introduced, the accuracy rises to 92.36%. This evaluation encompassed

a total of 10 subjects. Importantly, the proposed fusion technique led to a substantial enhancement in the accuracies across all word categories. The results presented in Tables 3.5 to 3.11 consistently demonstrate that SBERT outperforms USE in the classification of spoken words, particularly when utilizing phoneme-based text transcripts, both with and without stress markers.

- A similar trend is observed for distinct speakers, where the image-based classification approach attains an accuracy rating of 98.02% for spoken word classification among 35 subjects. Once again, the proposed fusion technique significantly improves the accuracies across all word categories.
- For the 35-word subjects, the amalgamation of (vowels + plosives) and (vowels + fricatives) yields a test accuracy of 74.82% without stress markers in the case of SBERT. The same combination achieves a test accuracy of 75.35% without stress markers for SBERT in the context of the 35-word subjects.
- The combination of (vowels + plosives) and (vowels + fricatives) for the 10-word subjects exhibits a test accuracy of 84.72% in the absence of stress markers, with embeddings sized at 512 × 1. Achieving a test accuracy of 85.01% in the 10-word subjects, the combination of (vowels + plosives) and (vowels + fricatives) demonstrates effectiveness, particularly in the presence of stress markers, with embeddings sized at 512 × 1.
- Achieving commendable test accuracy, the amalgamation of (vowels + plosives) and (vowels + fricatives) for the 10-word subjects proves effective, particularly in the presence of stress markers, with embeddings sized at 768 × 1.When stress markers are introduced, the accuracy increases to 85.79% with embeddings sized at 768 × 1. When stress markers are removed, the accuracy decreases to 85.13% with embeddings sized at 768 × 1. Therefore, we favour SBERT over USE for achieving superior classification scores.
- The ablation analysis, conducted on phoneme embeddings (without stress markers) using the SBERT transformer, with embeddings sized at 768 × 1 per spoken word category among 10 subjects, yields the following results:
  - Combination of vowels and fricatives : 58.20%.
  - Combination of vowels and plosives : 79.29%.
  - Combination of vowels and lateral fricatives : 42.40%.
  - Combination of vowels and glides : 56.38%.

- Combination of vowels and nasals : 65.92%.
- Combination of vowels and taps : 63.23%.
- Combination of vowels and liquids : 63.74%.
- Combination of vowels and trills : 62.13%.

- The ablation analysis, conducted on phoneme embeddings (with stress markers) using the SBERT transformer, with embeddings sized at 768 × 1 per spoken word category among 10 subjects, yields the following results:
  - Combination of vowels and fricatives : 59.37%.
  - Combination of vowels and plosives : 81.08%.
  - Combination of vowels and lateral fricatives : 43.67%.
  - Combination of vowels and glides : 43.93%.
  - Combination of vowels and nasals : 66.29%.
  - Combination of vowels and taps : 64.54%.
  - Combination of vowels and liquids : 63.94%.
  - Combination of vowels and trills : 70.81%.
- It becomes evident that, within the scope of in-depth linguistic analysis, stress markers play a pivotal role and cannot be disregarded, given their significance in identifying spoken words using text transcripts. It's worth noting that, contrary to concerns about high dimensionality leading to the computational challenges, our findings indicate that high dimensionality contributes to more accurate results. In our specific context, the Sentence-BERT transformer slightly outperforms the USE.
- Utilizing only nasals and glides yields improved results in the absence of suprasegmental information, while the inclusion of suprasegmental information enhances the performance of other phonological attributes.

Despite the implementation of our proposed technique, notable improvements in accuracy for specific word categories remain elusive. Further exploration and refinement are necessary to effectively enhance recognition performance in these specific categories. Additional research and experimentation may be required to address the challenges and achieve significant accuracy improvements in those word categories. In Tables 3.5 - 3.11, we conducted a comprehensive exploration of various phonological and spectrogram assessments, with a specific focus on vowels and plosives. Vowels, characterized by their production with minimal or no constriction of the vocal tract, allowing the unrestricted

airflow through the mouth, hold a central position in forming the core of syllables across many languages. Their recognition and proper articulation are crucial for speech intelligibility.

Table 3.5 Enhancing Speech Recognition with Phonological Stress Markers: A Sentence-BERT Based Evaluation on Google Speech Command Dataset's 10-Word Category

LINGUISTIC PHONOLOGICAL APPROACH (With stress marker) (768 X1) Sentence-BERT	TRAIN ACCURACY (%)	VALIDATION ACCURACY (%)	TEST ACCURACY (%)
Vowels + Plosives	81.54	81.38	81.08
Vowels + Fricatives	60.09	59.69	59.37
Vowels + Lateral Fricatives	44.68	43.28	43.67
Vowels + Glides	44.44	44.34	43.93
Vowels + Nasals	67.67	67.56	66.29
Vowels + Traps	65.55	64.84	64.54
Vowels + Liquids	64.92	64.79	63.94
Vowels + Trills	72.21	70.90	70.81

Table 3.6 Enhancing Speech Recognition without Phonological Stress Markers: A Sentence-BERT Based Evaluation on Google Speech Command Dataset's 10-Word Category

LINGUISTIC PHONOLOGICAL APPROACH (Without stress marker) (768 X1) Sentence-BERT	TRAIN ACCURACY (%)	VALIDATION ACCURACY (%)	TEST ACCURACY (%)
Vowels + Plosives	81.13	79.90	79.26
Vowels + Fricatives	59.02	58.49	58.20
Vowels + Lateral Fricatives	44.75	43.75	42.40
Vowels + Glides	56.80	56.43	56.38
Vowels + Nasals	66.68	66.47	65.92
Vowels + Traps	65.74	64.15	63.23
Vowels + Liquids	64.69	64.67	63.74
Vowels + Trills	63.28	62.26	62.13

Table 3.7 Enhancing Speech Recognition without Phonological Stress Markers: USE-based Evaluation on Google Speech Command Dataset's 10-Word Category

LINGUISTIC PHONOLOGICAL APPROACH (Without stress marker) (512 X 1) USE embeddings	TRAIN ACCURACY (%)	VALIDATION ACCURACY (%)	TEST ACCURACY (%)
Vowels + Plosives	71.98	71.48	70.66
Vowels + Fricatives	58.82	58.81	58.23
Vowels + Lateral Fricatives	43.39	43.28	43.00
Vowels + Glides	57.14	57.06	55.51
Vowels + Nasals	67.69	66.69	66.55
Vowels + Traps	65.39	65.38	65.37
Vowels + Liquids	64.87	64.77	64.64
Vowels + Trills	70.91	70.83	70.64

Table 3.8 Enhancing Speech Recognition with Phonological Stress Markers: USE-based Evaluation on Google Speech Command Dataset's 10-Word Category

LINGUISTIC PHONOLOGICAL APPROACH (With stress marker) (512 X 1) USE embeddings	TRAIN ACCURACY (%)	VALIDATION ACCURACY (%)	TEST ACCURACY (%)
Vowels + Plosives	80.31	80.23	79.54
Vowels + Fricatives	60.98	60.83	60.19
Vowels + Lateral Fricatives	44.54	44.45	43.39
Vowels + Glides	55.89	55.32	55.20
Vowels + Nasals	53.95	53.72	53.01
Vowels + Traps	65.64	65.44	65.31
Vowels + Liquids	65.17	64.98	64.61
Vowels + Trills	72.69	71.13	72.03

Table 3.9 Additional evaluation metrics are considered for the 10-Word Category of the Google Speech Command Dataset, specifically without the inclusion of phonological stress markers

APPROACH	Methodology (Without stress markers) S-BERT	TRAIN ACCURACY (%)	VALIDATION ACCURACY (%)	TEST ACCURACY (%)
Phonetic approach	Phonemes	91.28	91.19	90.64
Phonetic approach	VP + VF	85.83	85.42	85.13

Table 3.10 Some more evaluation measures on with phonological stress markers and spectrograms for Google Speech Command Dataset's 10-Word Category

APPROACH	Methodology proposed on 10-word categories	TRAIN ACCURACY (%)	VALIDATION ACCURACY (%)	TEST ACCURACY (%)
Phonetic approach	Phonemes (With stress markers) S-BERT	93.40	93.14	92.36
Phonetic approach	VP + VF (With stress markers) S-BERT	88.01	85.88	85.79
Visual-based approach	Tiny swin	99.85	99.76	98.92
Decision-level Fusion	Phonemes (With stress markers) S-BERT + Tiny swin	–	–	99.87

Table 3.11 Performance evaluation measures on phonological stress markers (with and w/o) and spectrograms for Google Speech Command Dataset's 35-Word Category

APPROACH	35-WORD CATEGORIES	TRAIN ACCURACY	VALIDATION ACCURACY	TEST ACCURACY
Phonetic approach	35-combo of VP + VF (Without stress markers) S-BERT	75.30%	75.00%	74.82%
Phonetic approach	35-combo of VP + VF (With stress markers) S-BERT	76.89%	75.40%	75.35%
Phonetic approach	35-combo of VP + VT (Without stress markers) S-BERT	77.45%	77.33%	76.98%
Phonetic approach	35-combo of VP + VT (With stress markers) S-BERT	78.99%	78.63%	77.69%
Phonetic approach	35-combo of VP + VN (Without stress markers) S-BERT	67.83%	66.72%	66.51%
Phonetic approach	35-combo of VP + VN (With stress markers) S-BERT	69.88%	69.81%	68.98%
Phonetic approach	35-combo of VP + VF (Without stress markers) S-BERT	77.81%	76.68%	76.66%
Phonetic approach	35-combo of VP + VF (With stress markers) S-BERT	76.96%	76.91%	76.25%
Visual-based approach	35 swin	98.10%	98.09%	98.02%
Phonetic approach + Visual-based approach (Spectrogram with tiny Swin)	Phoneme + Spectrogram	–	–	98.53%

Additionally, we adopted a linguistic phonological approach incorporating stress markers denoted as 0 (no stress), 1 (primary stress), and 2 (secondary stress). Stress levels wield significant influence over spoken utterances and can substantially affect the accuracy of SI text transcription methods. Our analysis underscores the significance of incorporating stress markers, as they contribute to remarkable accuracy levels in transcription, particularly when dealing with vowels and plosives. Specifically, when assessing the test accuracy of phonemes involving vowels and plosives, our findings indicate that integrating stress markers using Sentence BERT yielded an impressive accuracy of 81.08%, while the USE achieved 79.54%. However, without the inclusion of stress markers, the accuracy exhibited a slight decline to 79.26% for Sentence BERT and 70.66% for the USE. Notably, Sentence BERT outperformed the USE, particularly in the context of plosives, signifying its superiority in this domain. Furthermore, it would be beneficial to explore how diverse stress patterns in different languages might impact transcription accuracy and identify any language-specific adaptations that could enhance performance. Continual refinement and expansion of our phonological assessments have the potential to advance the field of speech recognition and text transcription, creating opportunities for broader applications in NLP and HCI. The tables 3.5 - 3.11 unequivocally depict the variable accuracies observed across different phonological attributes. Notably, the consistent augmentation of accuracy associated with incorporating stress markers underscores the pivotal role of stress in speech recognition and its potential impact on transcription accuracy. The table 3.11 provides an overview of the performance of various phonetic and visual-based approaches concerning the 35-word categories. The visual-based approach utilizing the Swin T transformer exhibits superior accuracy, nearly achieving perfection in results. Nevertheless, the phonetic approaches still showcase commendable accuracy levels, particularly when integrating stress markers, which prove instrumental in performance enhancement. Upon comparison with other SOTA techniques, our proposed approach clearly stands out as the top-performing method. We have demonstrated superior results and attained the highest accuracy among competing approaches. While specific word categories may offer room for improvement, our overall performance surpasses the current SOTA, accentuating the effectiveness and potential of our approach in advancing in the field of speech recognition. In our study, we delved into diverse phonological combinations using phoneme slicing. Upon analysing the GSCD for the 10-word categories, we achieved impressive training accuracy of 93.40% and test accuracy of 92.36%. Among various speech sounds, plosives, nasals, and trills played

significant roles in speech recognition. However, the combination of multiple distinct phonemes did not contribute significantly, whereas phonemes with stress markers proved to be valuable. Throughout our experiments, Sentence BERT consistently demonstrated high performance across various phoneme variations. While the judicious application of phoneme slicing can lead to improved results, it is imperative not to underestimate the importance of audio-based features in attaining enhanced accuracy. Our findings underscore the pivotal role of specific phonological attributes, especially stress markers, in enhancing speech recognition performance. However, it is equally important to recognize that combining diverse phonemes may not consistently result in substantial improvements. To advance our research further, it is crucial to continue exploring and optimizing phoneme slicing techniques while also investigating additional audio-based features and their potential synergies with phonological attributes. Striking the right balance between these elements holds the promise of unlocking even more robust and accurate speech recognition systems. The integration of both phonological and audio-based features offers the potential for superior results in speech recognition tasks. The ROC curves depicted in Figure 3.11 and Figure 3.12 provide a multiclass evaluation for classifying the 10-word categories using phonology with stress markers (0, 1, and 2). This evaluation encompasses various speech sounds, such as lateral fricatives, nasals, fricatives, glides, plosives, traps, liquids, and trills. Two types of embeddings, namely the 512 USE embeddings and the 768 sentence BERT embeddings, are employed for this analysis. Upon examining the results, it becomes evident that certain categories, particularly those with a substantial nasal component, are accurately identified when the nasal phoneme is applied. For instance, the category “NO” exhibits high accuracy in classification. Similarly, other speech sounds like fricatives, glides, and traps also exhibit promising performance in distinguishing specific categories.

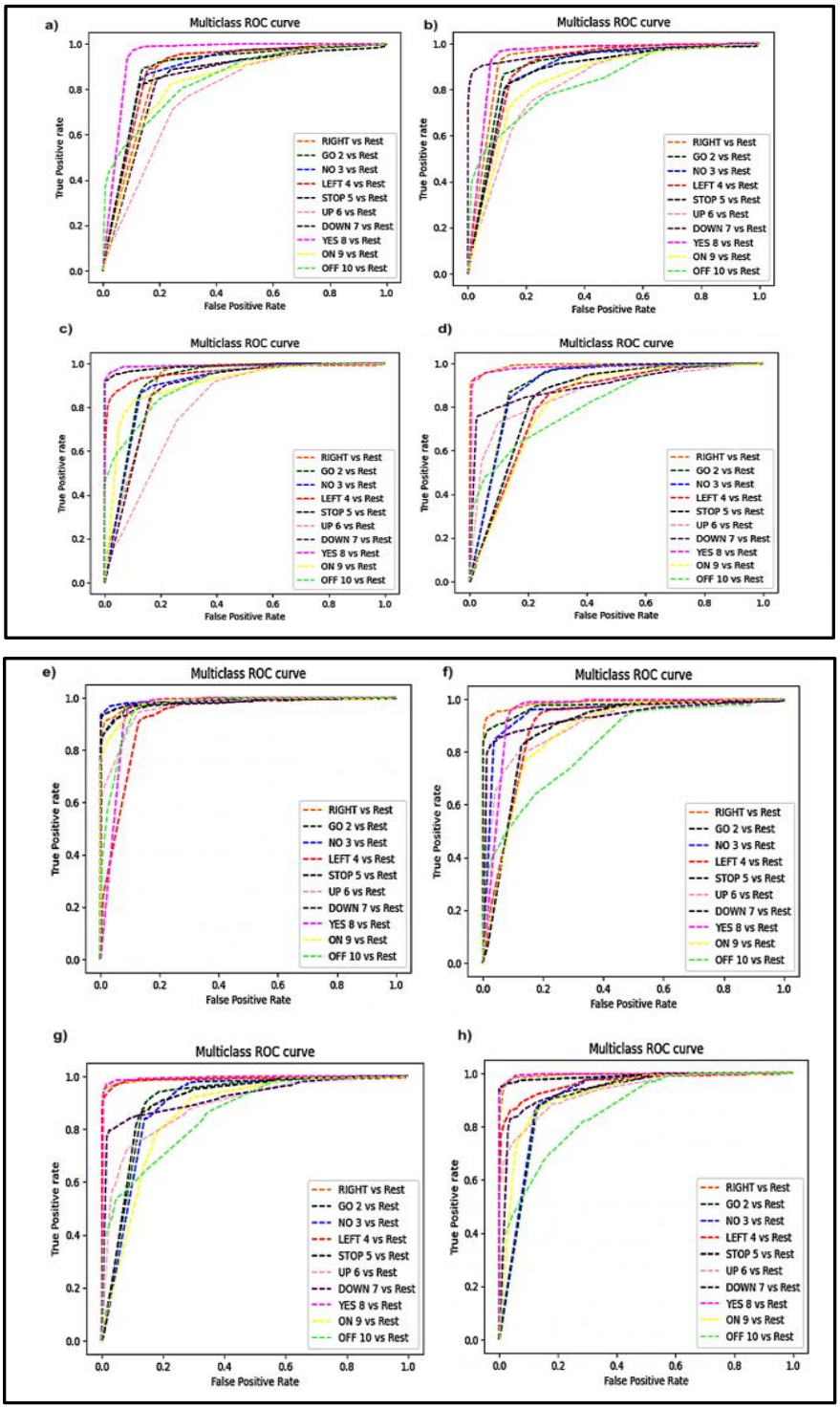


Figure 3.11. Multiclass evaluation ROC for classifying 10-word categories using phonology with stress markers (0,1, and 2) where a) lateral fricatives, b) nasals, c) fricatives, d) glides, e) plosives, f) traps, g) liquids and h) trills using 512 USE embeddings

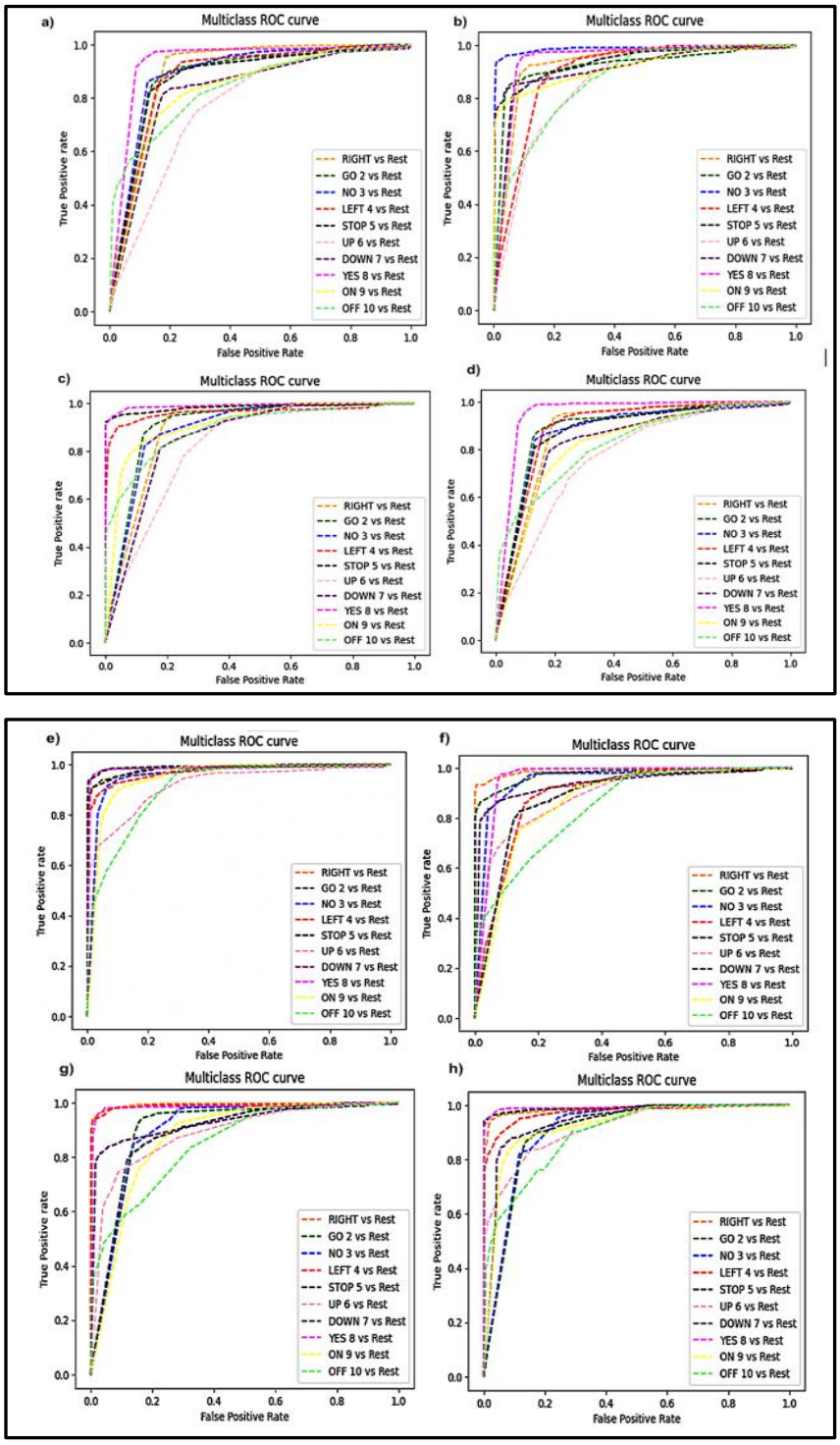


Figure 3.12. Multiclass evaluation ROC for classifying 10-word categories using phonology with stress markers (0,1, and 2) where a) lateral fricatives, b) nasals, c) fricatives, d) glides, e) plosives, f) traps, g) liquids and h) trills using 768 sentence-BERT embeddings

In summary, the insights gleaned from Figure 3.11 and Figure 3.12 underscore the significance of various speech sounds in phonological analysis and their profound influence on the multiclass classification of word categories. Remarkably, plosives emerge as the top-performing speech sound, consistently delivering the highest accuracy in classifying each category. Their robust representation and distinctive acoustic characteristics make them highly effective in discerning different word categories. Furthermore, the significance of trills should not be underestimated. Despite being relatively less explored, trills consistently demonstrate excellent performance for both USE embeddings and sentence BERT embeddings, underscoring the importance of considering trills as valuable phonological features in speech recognition tasks. To further enhance the outcomes, researchers should persist in exploring the unique contributions of different speech sounds and optimizing their integration into phonological analysis. Moreover, investigating how the amalgamation of multiple phonological features can bolster classification accuracy offers a promising avenue for future advancements. The outstanding performance of plosives and the notable role of trills underscore the importance of adopting a comprehensive approach to leverage a diverse set of phonological attributes for achieving superior results in the area of speech recognition.

The experimental outcomes clearly demonstrated substantial enhancements in voice recognition accuracy when compared to existing methodologies. Our system not only outperformed current approaches but also set a new benchmark in SWR. By incorporating linguistic insights and harnessing diverse resources, we achieved exceptional performance in the area of automatic speech recognition. Overall, this study constitutes a significant advancement in the field of automatic speech recognition, paving the way for innovative avenues in multimodal data analysis. The proposed approach represents a promising direction for future research endeavours within the domain of speech recognition, promising more sophisticated and efficient speech processing systems.

3.7 SIGNIFICANT OUTCOMES

Our study underscores the pivotal role of vowels and plosives in speech recognition, with stress markers significantly contributing to transcription accuracy. Focusing on these phonological features and exploring avenues for further improvement can drive advancements in speech recognition technologies and their versatile applications, catering

to a wide spectrum of users and languages. Trills, involving consonant sounds generated by rapid articulator vibrations, produced the second-highest accuracy among the various sounds analysed. When stress markers are integrated through Sentence BERT, the test accuracy with vowels and trills achieved 70.81%, while the USE scored 72.03%. In the absence of stress markers, accuracy dropped to 62.13% (Sentence BERT) and 70.64% USE.

Intriguingly, stress markers contributed to enhance accuracy, and the USE outperformed Sentence BERT in this particular context. To optimize results, further research should delve into trill variations across different languages and refine the incorporation of stress markers for improved speech recognition in diverse linguistic contexts. Nasals, encompassing consonant sounds created by allowing air to pass through the nose while obstructing the oral cavity, achieved noteworthy accuracy among the assessed sounds. When stress markers are included through Sentence BERT, the test accuracy with vowels and nasals is 66.29%, whereas the USE achieved 53.01%. Without stress markers, the accuracy remained relatively high at 65.92% (Sentence BERT), and 66.55% USE.

Surprisingly, despite stress markers generally improving accuracy, the USE outperformed Sentence BERT in this instance. Remarkably, the USE's embeddings, generated without considering stress markers, proved to be the most accurate.

To enhance performance, future investigations should delve into the distinctive aspects of nasals in various languages and optimize the utilization of stress markers to refine speech recognition across diverse linguistic contexts.

The methodology outlined above is motivated by the work discussed in the forthcoming section.

Living organisms communicate through speech, a process that involves the intricate analysis of spoken language to identify words and sentences. However, the pervasive influence of background noise presents a persistent challenge in achieving optimal speech recognition rates.

The current inadequacy in detection rates under noisy conditions necessitates dedicated research and potential interventions in the realm of speech recognition. To ameliorate the impact of background noise on speech recognition, this study proposes a novel approach employing a combination of median filtering and adaptive filtering. The methodology for speech command recognition involves a sequence of five key steps: first, the enhancement of signals through two parallel and independent speech enhancement models employing median and adaptive filtering; second, the extraction of 2D Mel spectrogram images from the enhanced signals; and third, the utilization of the tiny Swin Transformer for classification based on the obtained spectrogram images. The classification task involves the extensive ImageNet dataset, comprising 14 million images and approximately 150 GB in size. This study establishes the efficacy of decision-level fusion utilizing an audio-visual pair for robust speech recognition in the presence of background noise. The subsequent sections delineate the essential components of our investigation. In Section 3.8, we introduce a novel dense architecture designed for merging posterior scores. The experimental design and key findings are elucidated in Section 3.9, providing a cohesive overview of our empirical approach. The impetus behind engaging in speech recognition research and development stems from the overarching goal of elevating communication, expanding accessibility, optimizing operational efficiency, and fostering innovation across a spectrum of industries and applications.

3.8 PROPOSED METHODOLOGY

In this section, we delve into decision-level fusion (Mehra & Susan, 2021), leveraging average weighting scores for amalgamating results from two parallel channels. Past research (Mehra & Susan, 2021; Zhang et al. 2023) has underscored the effectiveness of late fusion in achieving superior categorization through the integration of features extracted from diverse segments of the enhanced input signal.

Post the learning phase, late fusion has been widely adopted by researchers to amalgamate multiple modalities (Mehra & Susan, 2021) into a cohesive representation (Das & Singh, 2023; Zhu et al. 2023). Through the concatenation of probabilistic scores, late fusion is anticipated to enhance the overall performance of the speech recognition system (Mehra & Susan, 2022). In Section 3.8.1, we expound upon the speech denoising process, while Section 3.8.2 delineates the application of the Tiny Swin Transformer for speech classification utilizing Mel spectrograms.

3.8.1 SPEECH DENOISING

Our proposed methodology integrates two distinct channels: the median filter and the adaptive filter (Rabiner & Wood, 2007; Gonzalez, 2007), as illustrated in Figure 3.13. The speech signals undergo processing through these filters, and the resultant signals are then utilized to derive two-dimensional Mel spectrograms. Both the median and adaptive filters dynamically adjust their parameters in response to the input signal, enabling them to effectively handle diverse types of noise and signal variations. These filters exhibit a selective noise removal capability while preserving essential signal features, effectively suppressing impulsive noise without distorting the underlying signal. This adaptability renders them well-suited for accommodating varying noise levels. In the context of speech processing, the median filter proves to be instrumental in reducing various types of noise while preserving the integrity of the signal and crucial edge information. Notably, its computational efficiency, ease of implementation, and suitability for real-time applications enhance its practical utility. The implementation of both median and adaptive filtering approaches leverages the SciPy Python library. Following the extraction of filtered speech and its conversion into 2-dimensional Mel spectrogram images using the librosa library, the subsequent step involves feeding these images into a pre-trained Tiny Swin Transformer network for image classification. Further elaboration on this process is provided in the ensuing section.

3.8.2 PRE-TRAINED SWIN-TINY TRANSFORMER MODEL

The filtered speech undergoes processing through two channels: a median filter and an adaptive filter. Subsequently, the filtered audio signals are transformed into 2-dimensional Mel spectrogram images using the librosa library. These images then undergo pre-processing

steps, including translation, rotation, and resizing, ultimately achieving a uniform dimension of 256 × 256. The standardized images are input into the Tiny Swin-T version 2 model, consisting of sequential layers, Swin Transformer blocks, and patch merging, with a core comprising 6 Swin Transformer blocks.

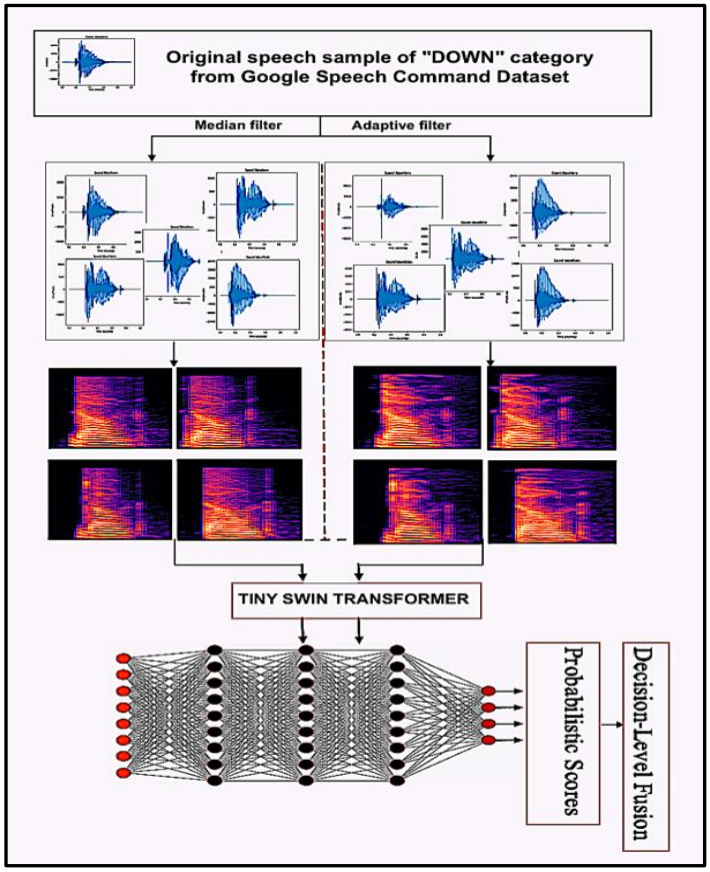


Figure 3.13. Workflow of Proposed Approach

The Swin Transformer, pre-trained on ImageNet (14 million images, 10,000 classes), stands as a benchmark in visual object recognition research. Renowned for its efficiency and accuracy, it surpasses the Vision Transformer (ViT) (Dosovitskiy et al. 2020). The Mel spectrograms are transformed into posterior scores through the Swin-Tiny Transformer, which, pre-trained on ImageNet, generates 10 posterior probabilistic scores for each category. The input features possess a dimension of 768. Our study employs a late fusion technique that combines posterior probabilistic scores from two distinct channels for speech command recognition, as depicted in Figure 3.13.

3.9 EXPERIMENTAL RESULTS

The experiments were conducted on Google Colab Pro++ using Python and GPU. The optimization utilized the Adam optimizer (Duchi et al. 2011) with a learning rate of 0.001, running for 100 epochs with a batch size of 32. The ReLU activation function introduced nonlinearity. In the speech-to-image process, embeddings were generated and input into a three-layered feed-forward network with unit sizes of 1024, 256, and 64, followed by a flattened layer, all employing ReLU activation. The Adam optimizer was chosen for optimization, and the model used sparse categorical cross-entropy as the loss function for multiclass classification. After obtaining classification scores, a decision-level fusion technique considered the average score value from two independent channels. Table 3.12 outlines the hyperparameters, while Table 3.3 compares our multimodal fusion approach with SOTA methods. Notably, our technique achieved a remarkable test accuracy of 99.85%, outperforming other methods across various domains, including keyword spotting, acoustic modeling, and attention-based encoder–decoder models. In summary, our proposed method demonstrated outstanding accuracy on the Google Speech Command dataset, surpassing SOTA approaches for categorizing the 10 speech command categories.

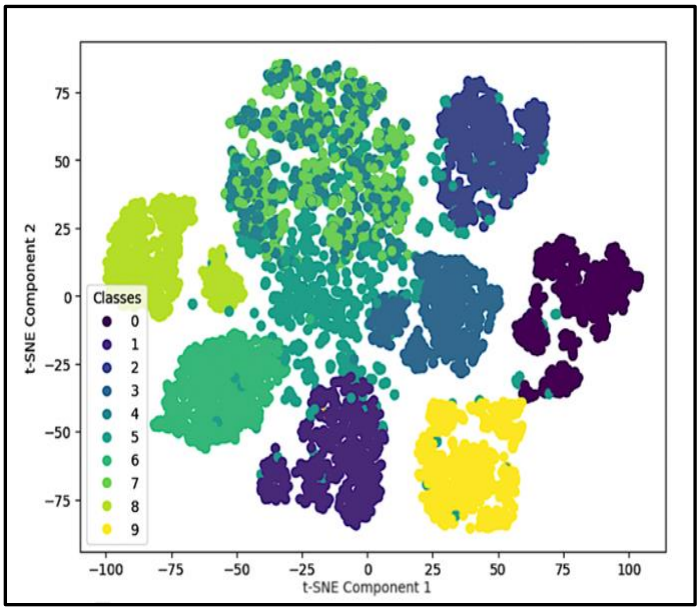


Figure 3.14. Multiclass classification evaluation with t-SNE for Speech Command Classification using adaptive filtering technique

The late fusion technique, combining outputs from the adaptive and median filtered channels, achieved a test accuracy of 99.85%, showcasing significant improvement over individual channels. This success can be attributed to the dataset's characteristics, encompassing SD samples with diverse accents and optimal data samples per category. Figure 3.5 and Figure 3.14 displays the distinct separation of 10 speech command categories using t-SNE for median filters, while adaptive filters present challenges in visualization. Both Figures 3.5 and 3.14 shows the superior performance of the median filter compared to the adaptive filter.

The fusion of these techniques produces the optimal outcome, as evidenced by the confusion matrix in Figure 3.15.

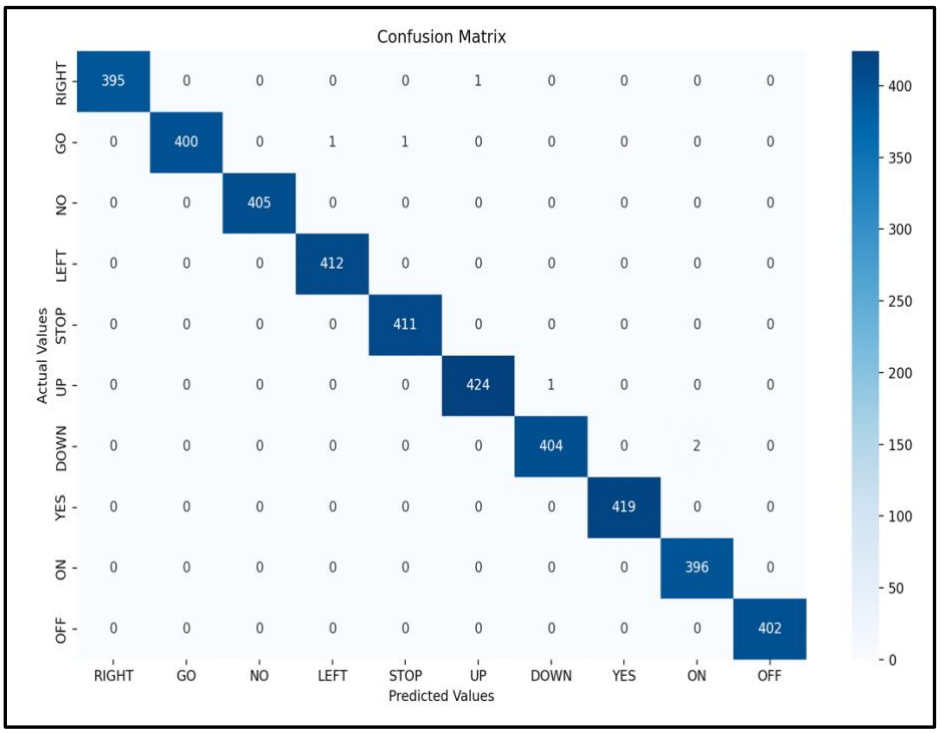


Figure 3.15. The confusion matrix is constructed to evaluate the fusion of deep frameworks adaptive and median filtering approaches

This amalgamation capitalizes on the strengths of two successful approaches, yielding advancements in classification performance beyond the SOTA benchmark. Table provides precision, recall, and F1-scores for each speech command category. Precision, a crucial metric, is defined as the ratio of true positives to the total count of positives identified by the model, as detailed in (3.1). This metric assesses the model's accuracy in recognizing positive instances while minimizing false positives.

Table 3.12 Hyperparameters of proposed approach

Parameters	Values
Dataset Size	1,21,30,79,830 bytes, 1.25 GB
Total Samples	38, 546
Testing Samples	4074
Epochs	100
Fully connected layer activation function	ReLU
Number of neural network layers	3

$$precision/\ positive\ predictive\ value\ (PPV) = \frac{tp}{[tp+fp]} \qquad (3.1)$$

Recall, also known as the true positive rate or sensitivity, is a crucial metric measured as the ratio of true positives to the total actual positives, as depicted in (3.2). This metric assesses the model's effectiveness in capturing all relevant positive instances, aiming to minimize false negatives.

$$sensitivity\ /\ recall = \frac{tp}{[tp + fn]} \qquad (3.2)$$

$$f1{-}Score = \frac{2 \times precision \times recall}{[precision+recall]} \qquad (3.3)$$

True positives, true negatives, false negatives, and false positives are represented by *tp*, *tn*, *fn*, and *fp*, respectively. The F1-score, a critical metric, is calculated as the harmonic mean of precision and recall, as shown in (3.3). This metric is essential for a comprehensive evaluation of the model's overall accuracy by considering both precision and recall simultaneously. In our experimental setup, we observed notably high precision, recall, and F1-score metrics, indicating robust classification performance. Table 3.13 presents a statistical analysis, affirming accurate classification across the majority of categories, with only minor deviations in a few classes. Table 3.13 details instances of misclassifications, revealing that specific word categories such as "NO," "LEFT," "STOP," "YES," "ON," and "OFF" exhibit zero misclassifications, highlighting the robustness of our method. Compared to cutting-edge approaches, our method distinguishes itself by achieving optimal classification performance and the highest accuracy. Figure 3.16 visually illustrates the tangible outcomes of our approach, demonstrating a test accuracy of 99.80%. The integration

of adaptive and median filtering-based probabilistic results through fusion elevates this accuracy to an impressive 99.85%. This enhancement underscores the efficacy of our approach in achieving superior classification performance through the synergistic utilization of filtering techniques.

Table 3.13 Statistical analysis per speech command categories

Categories	Precision	Recall	F1-score
Right	1.00	0.99	0.99
Go	0.99	1.00	0.99
No	1.00	0.99	1.00
Left	1.00	0.99	1.00
Stop	1.00	0.99	1.00
Up	1.00	0.99	0.99
Down	0.99	1.00	0.99
Yes	1.00	0.99	1.00
On	1.00	0.99	1.00
Off	1.00	0.99	1.00

3.10 SIGNIFICANT OUTCOMES

This thesis introduces an innovative approach to enhance speech command recognition by combining audio and image modalities through a late fusion technique. The methodology integrates a feed-forward neural network model with median and adaptive filtering methods to enhance audio signals. Utilizing the GSCD with 10-word categories, the approach achieves an impressive test accuracy of 99.85% through the integration of audio and image modalities. The fusion technique employs soft fusion, leveraging posterior class probabilities from two filtered channels extracted from each audio file. Feature extraction involves the Swin-Tiny Transformer, followed by a 3-layered feed-forward neural network. Compared to existing SOTA methods, the proposed fusion technique excels in classification accuracy, demonstrating effective capturing and utilization of information from both audio and image modalities for fine-grained speech command classification.

The integration of a pre-trained Swin-Tiny Transformer model, trained on an expansive image dataset, significantly contributes to the achieved accuracy. Notably, median filtering emerges as a superior pre-processing technique compared to adaptive filtering. The fusion of adaptive and median filtering-based probabilistic outcomes further enhances accuracy, resulting in an exceptional 99.85% success rate. One limitation is the increased memory demand, especially when applied to extensive datasets, a challenge shared across research endeavours on the same dataset.

Conducting Spectrogram and Analysis.

Given the intricate nature of dysarthric speech and the challenges it poses for comprehension, DSR has been proposed as a means to gauge intelligibility. However, its implementation requires extensive data and computational resources, rendering current techniques for objectively testing speech intelligibility laborious and somewhat arbitrary. Generic recognition systems often exhibit subpar performance in DSR. To address the complexities associated with speech impairment, this thesis presents a comprehensive ablation analysis of DSR across diverse speakers. Two distinct extractive transformer-based approaches are introduced for enhancing speech recognition. First, the use of Sepformer improves the speech signal, and the input of the enhanced audio signal is further processed by another transformer. Second, the Swin transformer is applied to log mel-spectrograms for image classification, pre-trained on 14 million annotated images from ImageNet. Pre-trained probabilistic scores obtained from both audio (SepFormer) and visual modalities (log mel-spectrogram) can be fine-tuned to classify spoken utterances. However, fine-tuning transformers necessitates considerable computational power and cost, making it impractical in the current scenario. To provide a cost-effective alternative, this thesis proposes a deep BiLSTM-GRU model for DSR. This model demonstrates outstanding performance on the EasyCall speech corpus, which encompasses cognitive characteristics. Remarkably, we achieved an accuracy of 98.56% for 20-word categories on dysarthric male speakers, 95.11% on dysarthric female speakers, and 97.55% on both dysarthric speakers, leveraging training output scores from audio-visual paired modalities trained on the proposed deep BiLSTM-GRU model.

Our approach demonstrates robust accuracy across various scenarios, outperforming other SOTA methods without the need for data augmentation.

The research presents several key contributions:

Our research introduces an optimized deep BiLSTM-GRU tailored for the classification of spoken utterances.

- Additionally, we propose a unified architecture for audio and visual pre-trained processing networks, employing two distinct transformers to capture essential features from both modalities.

- Through comprehensive analysis and testing on challenging dysarthric datasets, specifically the EasyCall corpus (Turrisi et al. 2021), we showcase the capability of our model to effectively replicate multimodal representations from descriptive audio and visuals. The results demonstrate ground-breaking advancements in Dysarthric Speech Utterance Classification (DSUC).

The section is organized as follows: Section 3.11, we delve into the details of our proposed approach. The intricacies of the experiments, encompassing the findings from the EasyCall corpus and the ablation study, are presented in Section 3.12. Finally, Section 3.13 encapsulates the concluding analysis of the research.

3.11 PROPOSED METHODOLOGY

Our innovative approach delves into the multimodal integration of audio and visual elements for DSUC. We introduce a sophisticated deep BiLSTM-GRU designed to classify dysarthric spoken utterances by leveraging salient features from both audio and visual domains. The subsequent sections outline the step-by-step procedural flow of our novel methodology.

3.11.1 DATA-HANDLING AND PRE-PROCESSING

In our innovative approach, we employed the Sepformer to elevate the quality of dysarthric audio samples. The Sepformer (Luo et al. 2020) encompasses an encoder, a decoder, and a masking network, all organized around a learned-domain masking approach. Within the masking network, two transformers operate (Dash & Solanki, 2020) within the dual-path processing block (Dash & Solanki, 2020), while the encoder, as detailed in (Nasersharif et al. 2023), embraces a fully convolutional design. The decoder reconstructs time-domain split signals based on the anticipated masks derived from the masking network. Ensuring reproducibility, the Sepformer is incorporated into the SpeechBrain toolkit. The Sepformer architecture showcases remarkable results in speech separation. Like other learned-encoder models (Ba et al. 2016), it employs short frames, a strategy supported by research for superior performance in such scenarios. We applied the Sepformer to each audio sample, resulting in enhanced audio samples.

Over the last decade, neural network advancements have significantly impacted speech augmentation and general audio source separation tasks. In contrast, traditional speech enhancement techniques analyze noise and clean speech spectra using statistical properties (Gerkmann & Vincent, 2018).

3.11.2 LOG-MEL SPECTROGRAMS (VOICEGRAMRS)

Following the acquisition of enhanced speech audio samples, we employ the Librosa package to generate log mel-spectrograms. This extraction process encompasses applying a short-time Fourier transform to the spoken signal, converting amplitude to decibels, and subsequently mapping frequencies onto the Mel scale. These 2D log mel-spectrogram representations, derived from audio samples featuring a diverse range of speakers, serve as inputs for speech classification in both the 10 and 20 spoken utterances of the EasyCall corpus. The refined log mel-spectrograms are then fed into the Swin transformer, as detailed in the subsequent section. Figure 3.16 illustrates a selection of extracted log mel-spectrograms derived from enhanced audio samples processed using Sepformer. These spectrograms depict the transformed audio data, demonstrating the effectiveness of

Sepformer in enhancing audio quality. Subsequently, the enhanced log mel-spectrograms are input into the Swin transformer, as detailed in the subsequent section.

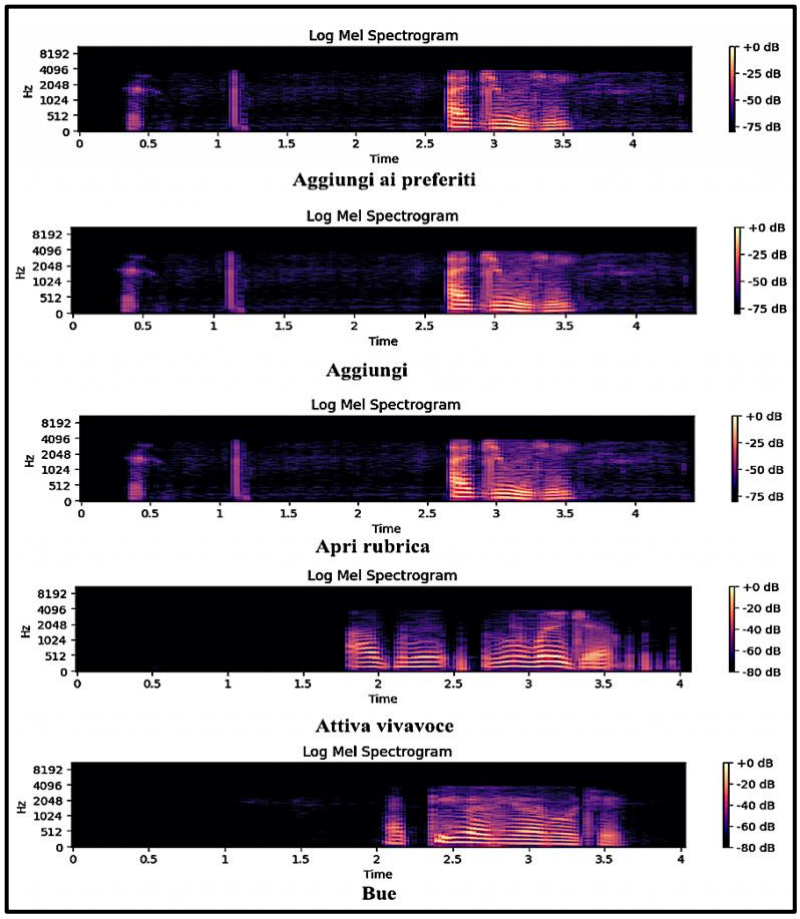


Figure 3.16. A snippet of a few samples of extracted log mel-spectrograms from enhanced audio samples

3.11.3 COMPUTER VISION SWIN TRANSFORMER

The Swin transformer incorporates crucial techniques to optimize its performance, including a residual-post-norm method with cosine attention to enhance training stability. It also adopts a log-spaced continuous position bias method and integrates SimMIM, a self-supervised pretraining approach, reducing the dependence on extensive labeled image datasets. In practical scenarios, the Swin transformer demonstrates remarkable efficacy in classifying log mel-spectrograms derived from 14 million pre-trained ImageNet annotated image samples. Its ability to accurately predict labels for provided images highlights its effectiveness in speech classification tasks. The Swin transformer's primary objective is to generate probabilistic scores for 20 spoken utterances by comparing log mel-spectrograms. Utilizing self-attention within a local window, it constructs hierarchical feature maps through

intricate computations. The output comprises these probabilistic scores, obtained by processing log mel-spectrogram data through a pre-trained network initially trained on Annotated ImageNet. It's noteworthy that we maintained the standard Swin model for this study, evaluating its impact on dysarthria data without modification. The pre-trained probabilistic scores from the Swin Transformer predict the class among 20 sentence categories by analyzing the log mel spectrogram from the Annotated ImageNet used during training. These scores offer insights into the likelihood of each sentence category. Our strategy aims to enhance speech recognition performance by leveraging salient features from both acoustic and visual modalities within a unified single-stream deep framework. This entails implementing a deep strategy that integrates a single-branch fusion of audio and visual modalities shown in Figure 3.17.

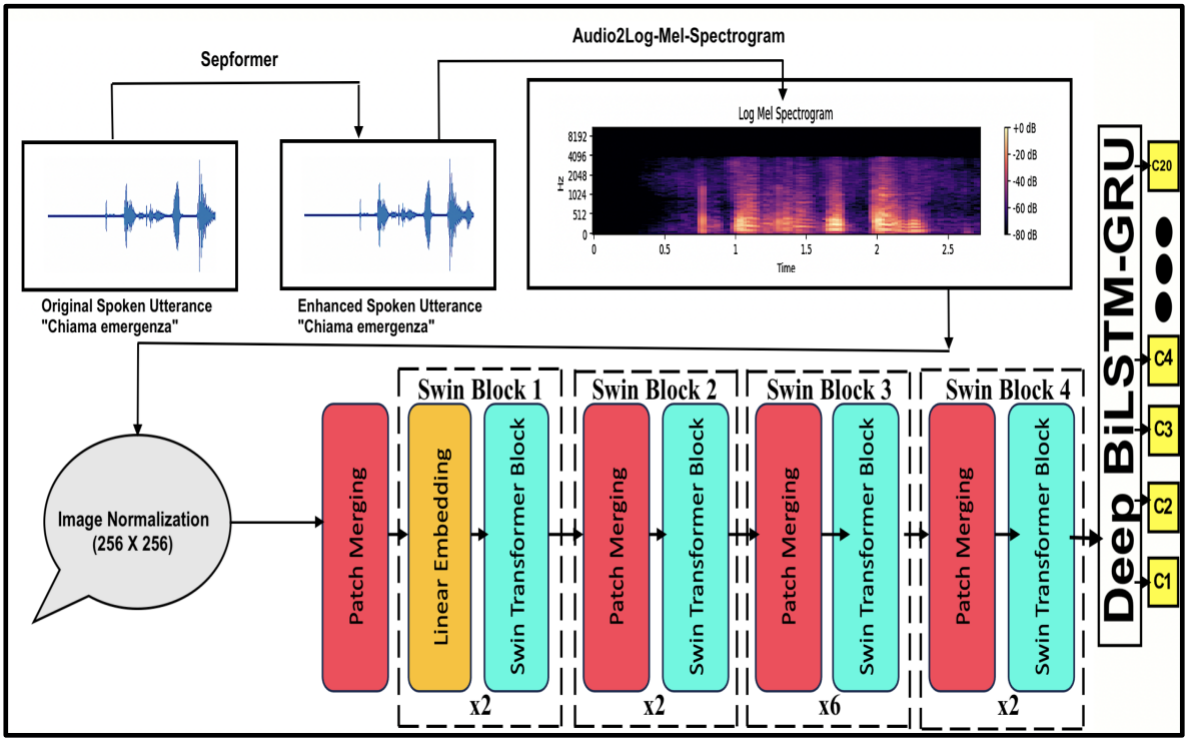


Figure 3.17. A step-by-step procedure for the proposed approach

3.11.4 Proposed Unified Deep Framework For Single-Branch Integration Of Audio And Visual Modalities

Our work is predicated on the fundamental assumption that the integration of audio and visual modalities within a unified single-stream deep framework significantly improve the performance of the speech recognition system. Figure 3.18 illustrates our deep BiLSTM-GRU model.

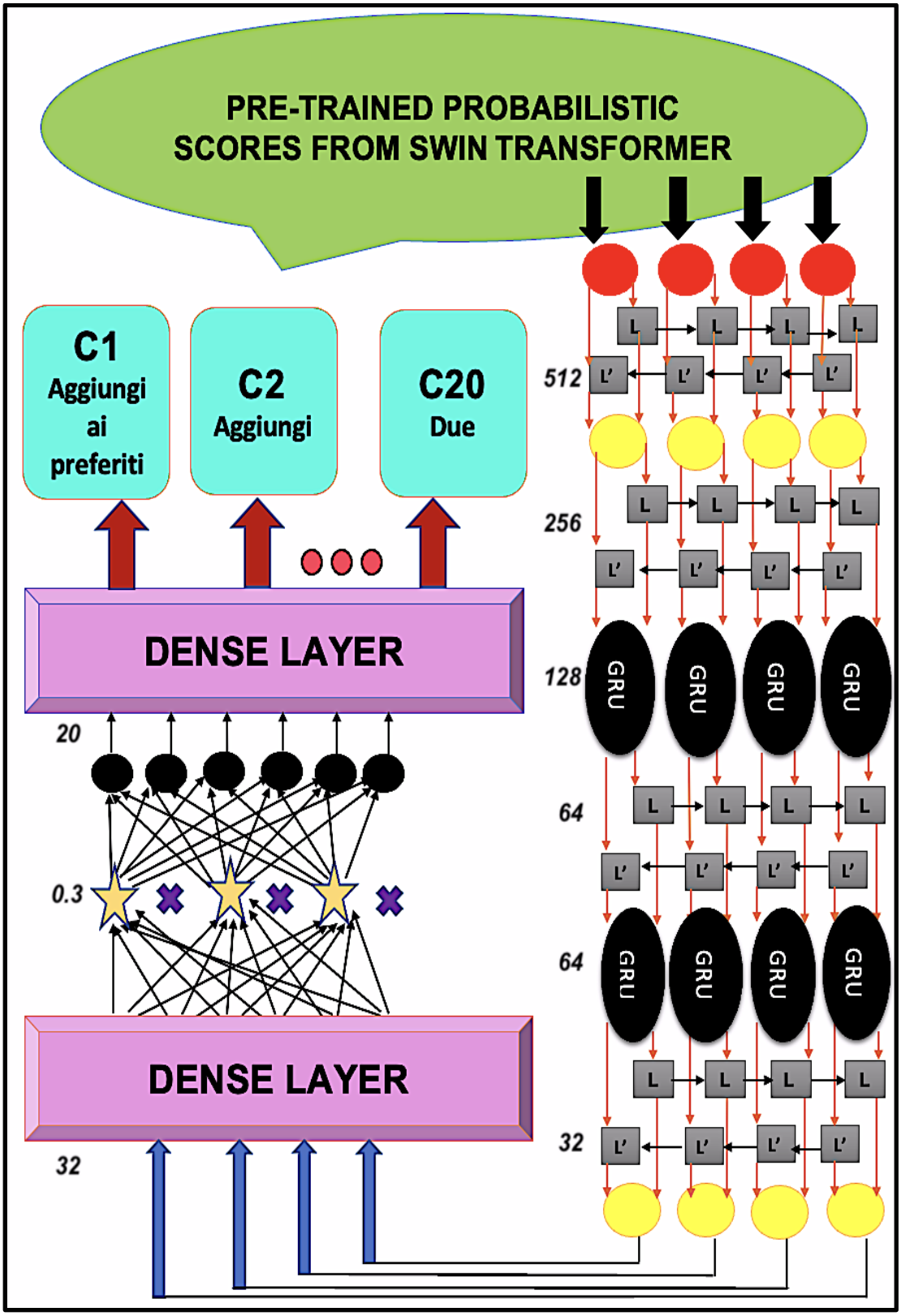


Figure 3.18. A proposed deep bi-LSTM-GRU model for dysarthric spoken utterance classification

To enhance classification accuracy, the BiLSTM-GRU is employed to process the output scores obtained from the Swin transformer. These scores are extracted from the Swin transformer and then input to the BiLSTM-GRU for further processing. The architecture of the deep BiLSTM-GRU model, depicted in Figure 3.18, includes a layer of BiLSTM with cells denoted by L for the forward LSTM and L' for the backward LSTM. The deep BiLSTM-GRU model processes the single-branch fusion of audio and visual modalities as inputs, computing them for each test sample through the utilization of the pre-trained model. Our proposed deep BiLSTM-GRU model features two initial BiLSTM layers with 512 and 256 units each, a gated recurrent layer with 128 units and the tanh activation function, a high-level combination of alternate BiLSTM and GRU layers with 64 and 32 units, along with dense 32 layers, dropout 32 layers, and dense 10 layers. This model is characterized by its depth, incorporating a total of 4 BiLSTM layers and 2 GRU layers. The forward and backward hidden states are obtained by feeding the input features into two LSTMs—one forward and one backward. The deep BiLSTM-GRU model utilizes the softmax activation function to generate probabilities for each class prediction. The primary assumption in our work is that merging acoustic and visual usage in single streams in a deep framework enhance the performance of the speech recognition system. To achieve this, we employ a single stream branching strategy that combines the strengths of both audio and visual modalities. The architecture of our proposed deep BiLSTM-GRU model, used for training audio and visual modalities together, is described. Basic RNNs face challenges in handling lengthy sequences, leading to information loss. To address this, LSTM models are introduced, featuring three gates—forget gate, input gate, and output gate—allowing the preservation of important information over extended sequences. The bi-directional LSTM processes the hidden states from both ends, contributing to the effective classification of sequential data such as audio and image data. In the LSTM equations (3.4-3.7), the input and previous hidden state combine to form the vector X, denoted as x_t, representing audio vectors derived from sequential data. The forget, input, and output gates are controlled by the biases b_i, b_f, and b_o, along with weight matrices W_i, W_f, and W_o, calculated during the training phase. The sigmoid function, denoted as σ, plays a key role in these calculations.

$$X = [h_{t-1}; x_t] \tag{3.4}$$

$$ft = \sigma (W_f X + b_f) \tag{3.5}$$

$$i_t = \sigma\,(W_i\,X + b_i) \tag{3.6}$$

$$o_t = \sigma\,(W_o\,X + b_o) \tag{3.7}$$

(3.5) to (3.7) represent gate activations with symbols *f*, *i*, and *o*. The hyperbolic tangent function is denoted as *tanh*, and the * signifies element-wise multiplication. In (3.8), the previous and current cell states are represented by c_{t-1} and c_t, respectively.

$$c_t = f_t * c_{t-1} + i_t * \tanh(W_c * X + b_c) \tag{3.8}$$

The hidden state at time-step t is represented as:

$$h_t = o_t * \tanh(c_t) \tag{3.9}$$

BiLSTM utilizes two hidden layers, one for the forward pass and another for the backward pass. The final hidden state results from concatenating the hidden states computed in both passes—forward ($\overrightarrow{ht}$) and backward ($\overleftarrow{ht}$). The effectiveness of BiLSTM-GRU is noteworthy, providing improved outcomes. The incorporation of GRU facilitates smoother training, enhancing overall training effectiveness. Leveraging the advantages of bidirectional circulation neural networks, this work employs BiLSTM-GRU, integrating advanced learning technology. In the neural network, LSTM is initially employed, processing the input x_t simultaneously with the previous node $\overrightarrow{ht}$. It undergoes the forget gate, input gate, and output gate to obtain the output htb. Subsequently, the output $\overleftarrow{ht}$ serves as the input for the preceding stage of GRU, producing the most recent state h_t through the update gate and reset gate. The model is inherently bidirectional, and the specific machine modifications involve inputting all generated eigenvalues into the model for training. After pre-processing life data from a few rolling bearings, parameters for LSTM and GRU are initialized. The output layer employs softmax for generating the probability distribution, while the RNN's hidden layer utilizes tanh as the activation function.

$$ht = \tanh(W_h * h_{t-1} + W_x * x_t) \tag{3.10}$$

The RNN input at time *t* is denoted as x_t. In the bidirectional RNN, the forward and reverse cells employ LSTM and GRU cells, respectively. The data computed in both directions is accumulated by these two cells to produce the combined result.

3.12 Experimental Results And Discussions

3.12.1 EasyCall Corpus: A Dysarthric Speech Dataset

A total of 31 individuals with dysarthria (11 females, 20 males) and 24 healthy speakers (10 females, 14 males) participated in providing utterances for the EasyCall corpus (Turrisi et al. 2021). Inclusion criteria for dysarthric speakers included age greater than or equal to 18 and dysarthria attributed to specific conditions such as Parkinson's disease, Huntington's disease, Amyotrophic Lateral Sclerosis, peripheral neuropathy, or myopathic or myasthenic lesions. Exclusion criteria encompassed aphasic syndromes, dementia, and intellectual incapacity. The dysarthria severity for each dysarthric speaker was assessed by an expert neurologist using the Therapy Outcome Measure (TOM), which assigns scores ranging from 1 to 5, with 1 indicating mild dysarthria, 2 for mild-moderate, 3 for moderate, 4 for moderate-severe, and 5 for severe dysarthria. The challenges associated with each word category are summarized in Table 3.14. Additionally, Figure 3.19 illustrates the distribution of training and testing samples in the dysarthric speech corpus.

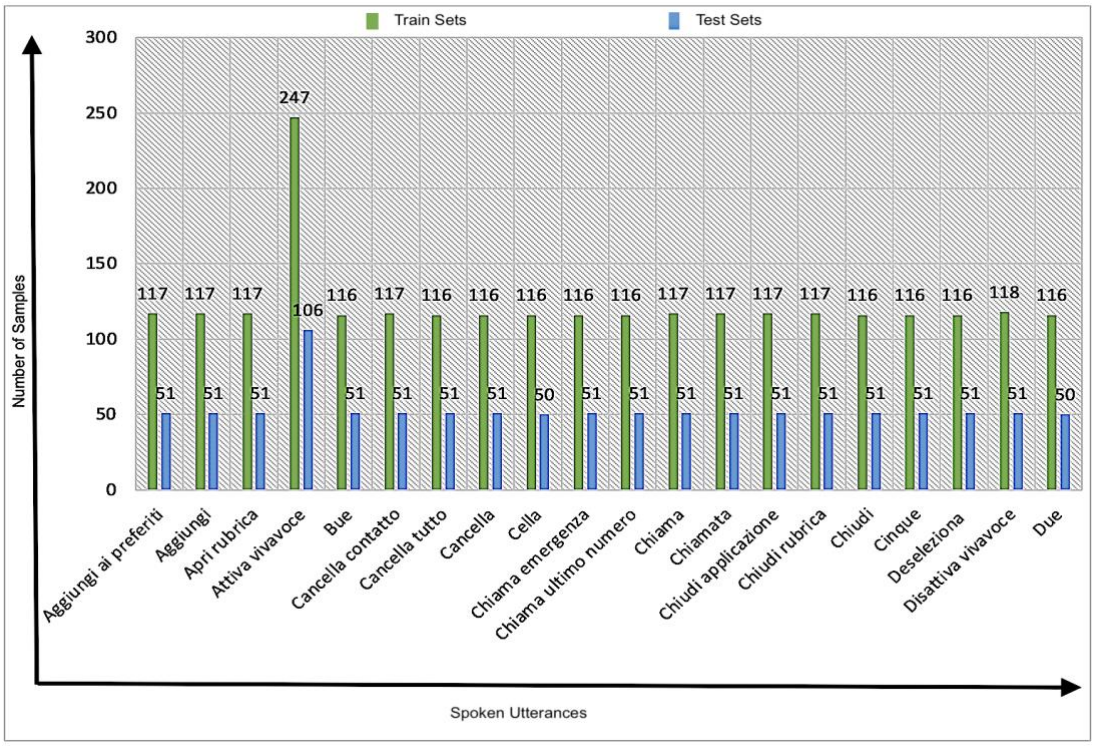


Figure 3.19. Characteristics of EasyCall Corpus for first 20 spoken words

Table 3.14 Summary of EasyCall Corpus data used in data pre-processing

Spoken Utterance	Speaker code	Type of dysarthria	Therapy Outcome Measure (TOM)	*Number of Sessions*	*Number of wav files*
Aggiungi ai preferiti	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	168
Aggiungi	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	168
Apri rubrica	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	168
Attiva vivavoce	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	353
Bue	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	167
Cancella contatto	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	168
Cancella tutto	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	167
Cancella	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	167
Cella	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	166
Chiama emergenza	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	167
Chiama ultimo numero	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	167
Chiama	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	168
Chiamata	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	168

Chiudi applicazione	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	168
Chiudi rubrica	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	168
Chiudi	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	167
Cinque	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	167
Deseleziona	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	167
Disattiva vivavoce	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	169
Due	*[F01-F11, M01-M20]*	*Paretic, cerebellar, extrapyramidal, pyramidal*	[_, 1, 2, 3, 4, 5]	[2, 3, 4, 5, 6]	166

3.12.2 EXPERIMENTAL DETAILS AND ANALYSIS

The experiments were conducted using Python version 3.10.0 on a Mac system running macOS Big Sur with an M1 chip. Due to the computational intensity associated with handling audio samples and transformers, we executed the experimental computations using the Librosa toolkit on Google Colab Pro, which is equipped with a GPU and provides 32 GB of RAM. Before applying the deep BiLSTM-GRU model to the pre-trained probabilistic scores, we computed the log mel-spectrogram feature matrix for each raw audio file, resulting in a 256 X 256 dimension. This matrix serves as input to the Swin transformer. Subsequently, the pre-trained probabilistic scores obtained from the Swin transformer are input into the trained deep BiLSTM-GRU model. For the ongoing experiment, we opted for 100 epochs. The model's nonlinearity is enhanced by the "ReLU" activation function. SGD is employed for regulating the learning rate, utilizing the Adam optimizer known for its reliable performance across various classification tasks. The Adam optimizer stores and utilizes the average of past gradients, contributing to the determination of current gradients and preserving the average of earlier gradients. Comparing our multimodal fusion technique to current best practices in Table 3.15, the results demonstrate our method's superior

performance in the fully-resourced data environment, achieving a maximum test accuracy of 97.64%.

Table 3.15 Performance analysis with different techniques by evaluation measures on EasyCall corpus

20 Spoken Utterances (Male + Female)	Macro	Accuracy (%)	Mathew's correlation coefficient	Cohen's kappa coefficient
MFCC + LSTM (Zia & Zahid, 2019)	0.4099	40.98%	0.4084	0.4084
MFCC + LSTM (Wazir et al. 2019)	0.4361	43.63%	0.4353	0.4353
Mel-Spectrogram + LSTM (Lezhenin et al. 2019)	0.3787	37.85%	0.3777	0.3777
GFCC + CNN (Abdelmaksoud et al. 2021)	0.4345	43.44%	0.4332	0.4332
MFCC + CNN (Haque et al. 2020)	0.4123	41.21%	0.4112	0.4112
MFCC + CNN (Kherdekar& Naik, 2021)	0.4325	43.21%	0.4315	0.4315
Hybrid CNN + BiLSTM (Passricha & Aggarwal, 2019)	0.4711	47.08%	0.4701	0.4701
Our proposed deep feature-level fusion bi-LSTM-GRU	0.9769	97.64%	0.9756	0.9756

Our proposed method, in contrast to CNN with input as a 2D matrix of MFCC features, outperforms in speech recognition tests. Even when compared to LSTM using Mel Spectrogram as input, our approach exhibits superior performance. The proposed approach surpasses the state of the art in classifying 10- and 20-word categories with dysarthric speech challenges in both male and female speakers of the EasyCall dataset, achieving a high accuracy of 97.64%. Despite the dataset's slight imbalance, the attained high accuracy is credited to the multimodal nature of the model, harnessing the synergies of two pre-trained models for the integration of audio and image information. The Matthews correlation coefficient (3.11) for the proposed approach is 0.9756, indicating a robust performance. Our approach demonstrates a Matthews correlation coefficient (MCC) of 0.9756, indicating a well-balanced assessment of classification performance, particularly beneficial for imbalanced datasets. MCC considers true positives, true negatives, false positives, and false negatives, offering a comprehensive measure ranging from -1 to 1, where 1 signifies perfect prediction, 0 implies no improvement over random prediction, and -1 indicates complete

disagreement between prediction and observation. The MCC score, closer to 1, signifies the superior performance of our model. Cohen's kappa coefficient serves as a robust metric for assessing the level of agreement among individuals when categorizing items, while also accounting for agreement that could occur by chance. The scale ranges from 0 to 1, with 1 indicating perfect agreement. Remarkably, our methodology yielded an exceptionally high score of 0.976, underscoring the substantial consensus among various raters or observers in the classification process.

$$\text{MCC} = (tp*tn - fp*fn) / \sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)} \quad (3.11)$$

$$\text{k} = \frac{po-pe}{1-pe} \quad (3.12)$$

The accuracy is calculated based on the values of true positives, true negatives, false positives, and false negatives as defined in (3.13).

$$Accuracy\ (\%) = (tp + tn)/(tp + fp + tn + fn) * 100 \quad (3.13)$$

The negative predicted value (NPV), false positive rate (FPR), false negative rate (FNR), and false discovery rate (FDR) for each spoken word category are detailed in Tables 3.16-3.18. In a symbolic representation, FPR, FNR, and FDR are denoted by (3.15), (3.16), and (3.17), respectively.

$$negative\ predictive\ value\ (NPV) = \frac{tn}{[fn + tn]} \quad (3.14)$$

$$false\ positive\ rate\ (FPR) = \frac{fp}{[fp + tn]} \quad (3.15)$$

$$false\ negative\ rate\ (FNR) = \frac{fn}{[fn + tp]} \quad (3.16)$$

$$false\ discovery\ rate\ (FDR) = \frac{fp}{[fp + tp]} \quad (3.17)$$

3.12.3 ABLATION STUDY

In this section, we conduct ablation research to investigate the impact of individual speakers (male or female) as well as their combination on our deep BiLSTM-GRU model. Within the proposed single stream branching framework, incorporating our deep BiLSTM-GRU, we assess the significance of each audio and visual component, pairwise combinations of male and female speakers, and the deep model itself. The ablation analysis results, summarized in Tables 3.16, 3.17, and 3.18, reveal the following insights. Table 3.19 presents the final

assessment for each speaker and class, emphasizing that combining audio and image features outperforms classification for the suggested deep framework. In our work, we have employed a single-branch approach for both audio and visual modalities, surpassing the performance of individually combining other widely used audio/visual components. Male speakers exhibit accuracy rates of 98.56%, female speakers at 95.11%, and both male and female speakers together at 97.64% (a total of 20 subjects). The proposed technique significantly enhances accuracies across all word categories, which are in Italian and spoken by dysarthria patients with mixed levels of speech impairment. The GRU module, positioned between the preceding BiLSTM layers and the alternate deeper-level BiLSTM and GRU layers, emerges as a crucial component of the proposed deep BiLSTM-GRU model. The comprehensive utilization of audio and visual information within a single branch is clearly linked to the enhanced accuracy in recognizing each spoken utterance. The success of our deep BiLSTM-GRU model, adept at learning from both audio and visual data for each modality, likely contributes to the observed high accuracy.

Table 3.16. Results of the classification using our proposed method for 20 spoken utterances by male speakers

Spoken Utterance	Sensitivity	Specificity	NPV	FPR	FNR	FDR	Precision
Aggiungi ai preferiti	0.9411	0.9984	0.9969	0.0015	0.0588	0.0303	0.9696
Aggiungi	0.9743	1.0000	0.9984	0.0000	0.0256	0.0000	1.0000
Apri rubrica	0.9655	1.0000	0.9984	0.0000	0.0344	0.0000	1.0000
Attiva vivavoce	0.9666	0.9968	0.9968	0.0031	0.0333	0.0333	0.9666
Bue	0.9459	0.9969	0.9969	0.0030	0.0540	0.0540	0.9459
Cancella contatto	1.0000	0.9984	1.0000	0.0015	0.0000	0.0312	0.9687
Cancella tutto	0.9130	1.0000	0.9970	0.0000	0.0869	0.0000	1.0000
Cancella	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Cella	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Chiama emergenza	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Chiama ultimo numero	0.9565	0.9970	0.9985	0.0029	0.0434	0.0833	0.9166
Chiama	1.0000	0.9984	1.0000	0.0015	0.0000	0.0285	0.9714
Chiamata	1.0000	0.9984	1.0000	0.0015	0.0000	0.0238	0.9761
Chiudi applicazione	0.9428	0.9969	0.9969	0.0030	0.0571	0.0571	0.9428
Chiudi rubrica	0.9411	0.9969	0.9969	0.0030	0.0588	0.0588	0.9411
Chiudi	0.9696	1.0000	0.9984	0.0000	0.0303	0.0000	1.0000
Cinque	1.0000	0.9969	1.0000	0.0030	0.0000	0.0476	0.9523

Deseleziona	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Disattiva vivavoce	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Due	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000

Table 3.17. Results of the classification using our proposed method for 20 spoken utterances by female speakers

Spoken Utterance	Sensitivity	Specificity	NPV	FPR	FNR	FDR	Precision
Aggiungi ai preferiti	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Aggiungi	0.9444	1.0000	0.9971	0.0000	0.0555	0.0000	1.0000
Apri rubrica	1.0000	0.9971	1.0000	0.0028	0.0000	0.0625	0.9375
Attiva vivavoce	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Bue	0.9500	1.0000	0.9971	0.0000	0.0500	0.0000	1.0000
Cancella contatto	0.8888	1.0000	0.9943	0.0000	0.1111	0.0000	1.0000
Cancella tutto	1.0000	0.9941	1.0000	0.0058	0.0000	0.0689	0.9310
Cancella	1.0000	0.9971	1.0000	0.0028	0.0000	0.0526	0.9476
Cella	0.9444	1.0000	0.9971	0.0000	0.0555	0.0000	1.0000
Chiama emergenza	0.8888	1.0000	0.9943	0.0000	0.1111	0.0000	1.0000
Chiama ultimo numero	1.0000	0.9916	1.0000	0.0083	0.0000	0.2500	0.7500
Chiama	0.9444	0.9971	0.9971	0.0028	0.0555	0.0555	0.9444
Chiamata	1.0000	0.9943	1.0000	0.0056	0.0000	0.1333	0.8666
Chiudi applicazione	0.9230	0.9971	0.9971	0.0028	0.0769	0.0769	0.9230
Chiudi rubrica	0.9583	0.9941	0.9970	0.0058	0.0416	0.0800	0.9200
Chiudi	0.8823	1.0000	0.9943	0.0000	0.1176	0.0000	1.0000
Cinque	0.8235	0.9914	0.9914	0.0085	0.1764	0.1764	0.8235
Deseleziona	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Disattiva vivavoce	0.8888	0.9971	0.9943	0.0028	0.1111	0.0588	0.9943
Due	0.9500	0.9971	0.9971	0.0028	0.0500	0.0500	0.9971

Table 3.18. Results of the classification using our proposed method for 20 spoken utterances, considering both male and female speakers

Spoken Utterance	Sensitivity	Specificity	NPV	FPR	FNR	FDR	Precision
Aggiungi ai preferiti	0.9705	1.0000	0.9984	0.0000	0.0294	0.0000	1.0000
Aggiungi	0.9743	1.0000	0.9984	0.0000	0.0256	0.0000	1.0000
Apri rubrica	0.9642	0.9984	0.9984	0.0015	0.0357	0.0357	0.9642
Attiva vivavoce	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Bue	1.0000	0.9984	1.0000	0.0015	0.0000	0.0270	0.9729
Cancella contatto	0.9393	0.9984	0.9969	0.0015	0.0606	0.0312	0.9687
Cancella tutto	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Cancella	0.9696	1.0000	0.9984	0.0000	0.0303	0.0000	1.0000
Cella	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Chiama emergenza	1.0000	0.9954	1.0000	0.0045	0.0000	0.0789	0.9210
Chiama ultimo numero	0.9600	1.0000	0.9985	0.0000	0.0400	0.0000	1.0000
Chiama	1.0000	0.9984	1.0000	0.0015	0.0000	0.0285	0.9714
Chiamata	0.9762	0.9984	0.9984	0.0015	0.0238	0.0238	0.9761
Chiudi applicazione	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Chiudi rubrica	0.9714	1.0000	0.9984	0.0000	0.0285	0.0000	1.0000
Chiudi	1.0000	0.9984	1.0000	0.0015	0.0000	0.0312	0.9687
Cinque	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Deseleziona	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	1.0000
Disattiva vivavoce	1.0000	0.9984	1.0000	0.0015	0.0000	0.0285	0.9714
Due	0.9714	1.0000	0.9984	0.0000	0.0285	0.0000	1.0000

Table 3.19. Accuracy per speakers per class

Spoken Utterance	20 spoken utterances of Male Speakers	20 spoken utterances of Female Speakers	20 spoken utterances of both Male and Female Speakers
Aggiungi ai preferiti	0.9985	1.0000	0.9956
Aggiungi	0.9985	0.9972	0.9985
Apri rubrica	0.9971	0.9972	0.9985
Attiva vivavoce	1.0000	1.0000	0.9942
Bue	0.9985	0.9972	0.9942
Cancella contatto	0.9956	0.9945	0.9985
Cancella tutto	1.0000	0.9945	0.9971
Cancella	0.9985	0.9972	1.0000
Cella	1.0000	0.9972	1.0000
Chiama emergenza	0.9957	0.9945	1.0000
Chiama ultimo numero	0.9985	0.9918	0.9956
Chiama	0.9985	0.9945	0.9985

Chiamata	0.9971	0.9945	0.9985
Chiudi applicazione	1.0000	0.9945	0.9942
Chiudi rubrica	0.9985	0.9918	0.9942
Chiudi	0.9985	0.9945	0.9985
Cinque	1.0000	0.9836	0.9971
Deseleziona	1.0000	1.0000	1.0000
Disattiva vivavoce	0.9985	0.9918	1.0000
Due	0.9985	0.9945	1.0000

In our research, we harnessed single-branch audio and visual modalities, surpassing the performance of combinations involving other popular individual audio/visual components. The findings revealed impressive accuracy ratings, with male speakers achieving 98.56%, female speakers reaching 95.11%, and an overall accuracy of 97.64% for both genders across 20 spoken utterances. The proposed technique notably elevated the accuracy levels for all utterances. Employing a 70:30 train-test ratio with 5-fold cross-validation, it's crucial to highlight that the dataset lacks a distinct folder demarcation for training and testing. Figures 3.20, 3.21, and 3.22 illustrate the confusion matrices of the proposed deep BiLSTM-GRU for the first 20 utterances of female speakers, male speakers, and both male and female speakers, incorporating a deep-single stream branching of audio and image modalities. The y-axis represents actual values, while the x-axis represents predicted values. The confusion matrix analysis reveals that the 20 utterances of male speakers exhibit exceptional performance, attributed to the enhancement of audio samples with SepFormer and the meta-learning of output scores from both SepFormer and Swin transformer. In the confusion matrix, "c1" and "c2" represent the predictions and actual classes of spoken utterances, respectively. "c1" stands for the first utterance on both axes, while "c2" represents the second utterance, and so on.

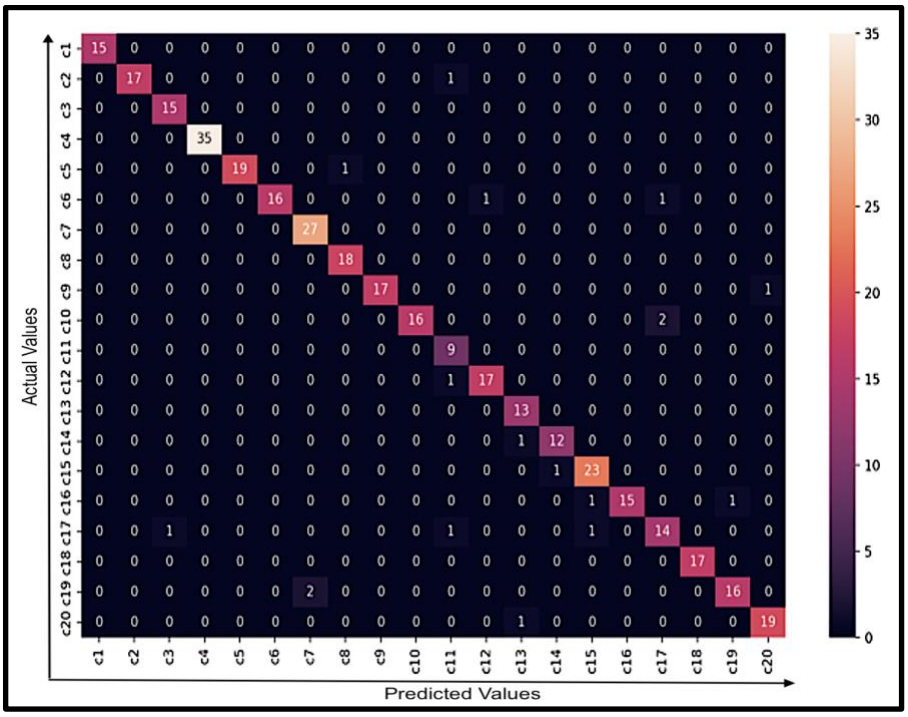


Figure 3.20. Confusion Matrix of first 20 instance of female speakers

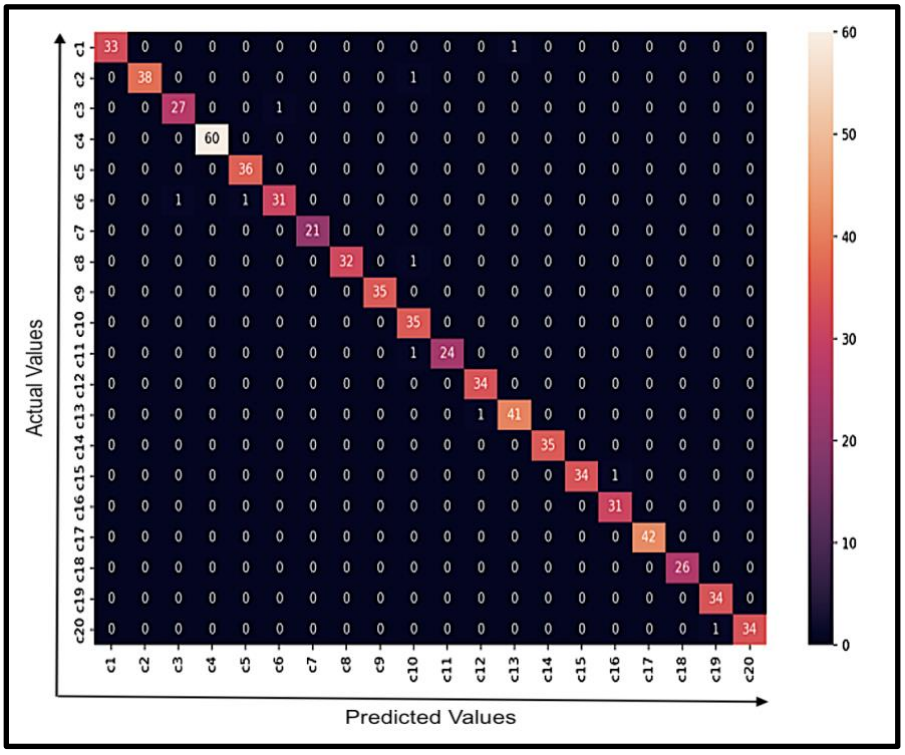


Figure 3.21. Confusion Matrix of first 20 instance of male speakers

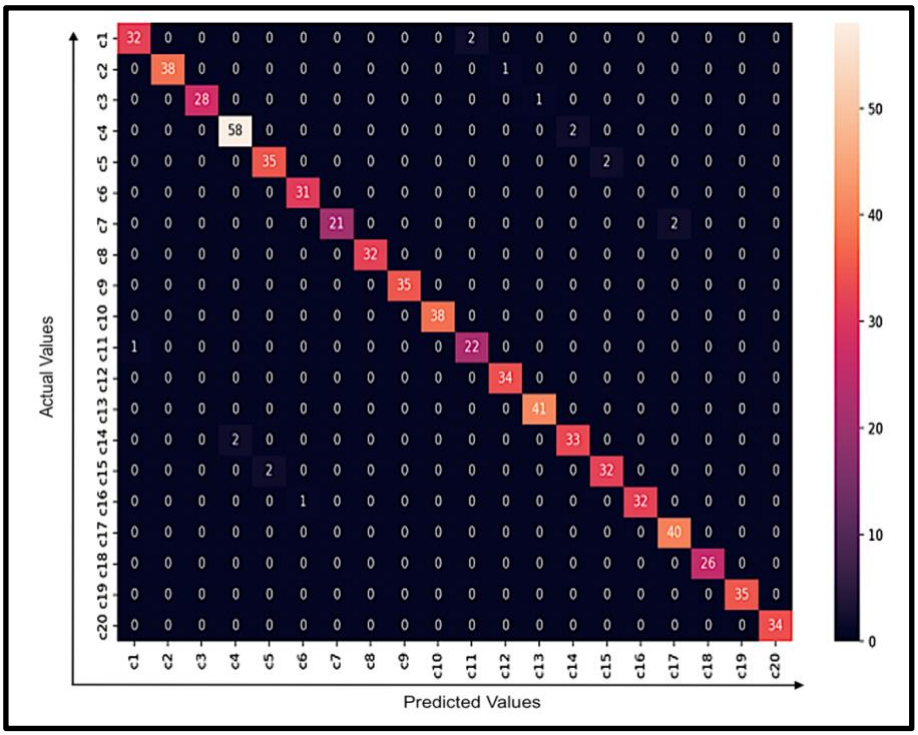


Figure 3.22. Confusion Matrix of first 20 instances of both gender speakers

The ROC curves in Figures 3.23, 3.24, and 3.25 for 10 subjects further demonstrate the effectiveness of our technique in accurately determining each speaker's utterance. Similarly, Figures 3.26, 3.27, and 3.28 for 20 subjects showcase the superior performance of our technique in speaker utterance accuracy. Even in the context of a challenging dataset, our approach proves effective in improving dysarthric speech recognition.

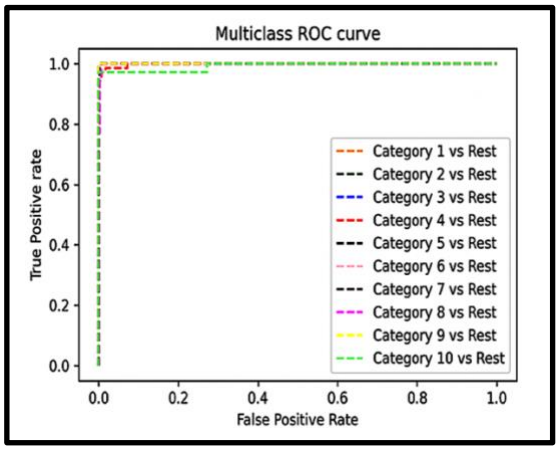


Figure 3.23. 10 spoken utterances of male

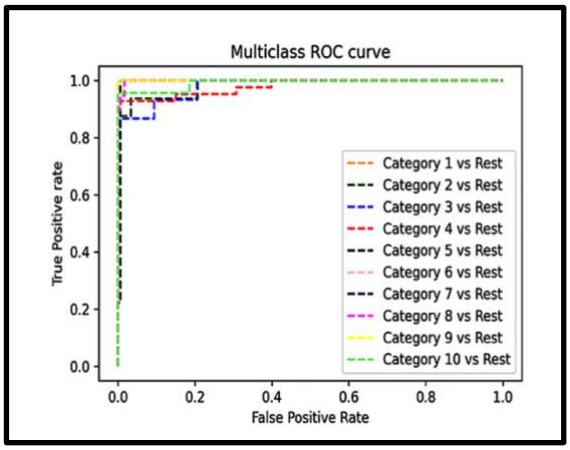


Figure 3.24. 10 spoken utterances of female

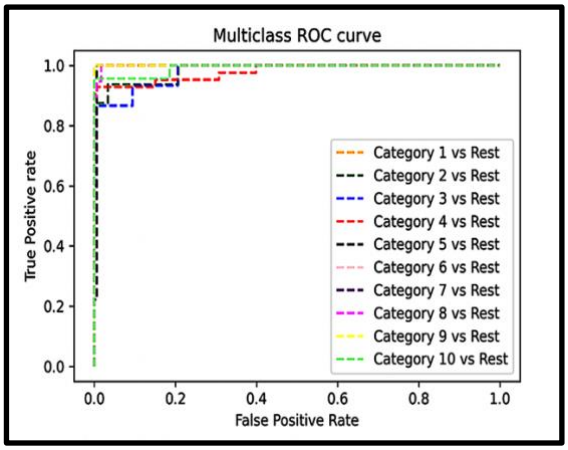


Figure 3.25. 10 spoken utterances of both

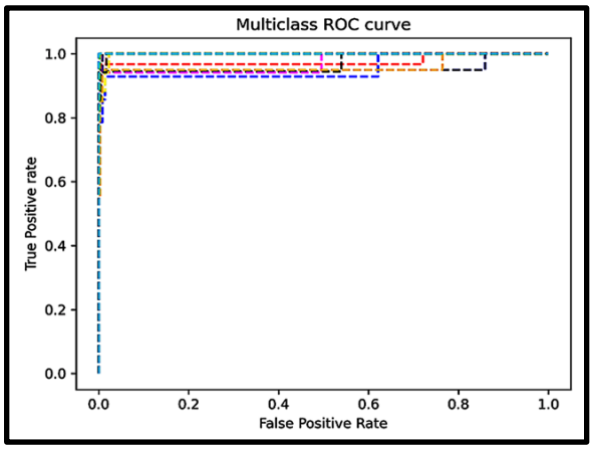


Figure 3.26. 20 spoken utterances of male

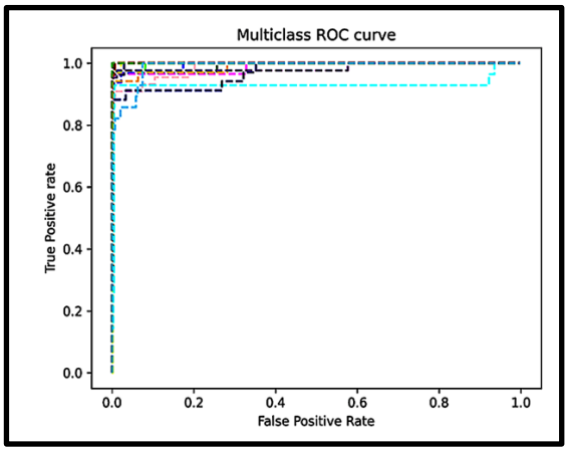


Figure 3.27. 20 speech utterances of female

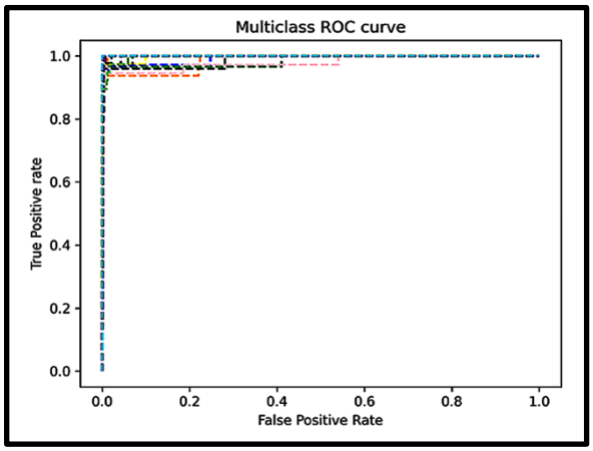


Figure 3.28. 20 speech utterances of both

It is important to acknowledge that dysarthric speech databases need to consider the possibility that individuals with speech impairments might use unconventional phrases or expressions due to difficulties in pronouncing specific words. Additionally, factors such as cultural and generational variations may contribute to differences in vocabulary between senior individuals, who are more susceptible to articulatory issues, and younger individuals.

3.13 Significant Outcomes

This objective introduces a detailed analysis of deep-salient features from audio and visual modalities using a novel deep BiLSTM-GRU model for individual modality learning. Leveraging training output scores from the deep BiLSTM-GRU model, our approach achieves high accuracies across various dysarthric speaker scenarios. Specifically, accuracy

rates include 98.56% for 20 utterances of dysarthric male speakers and 95.11% for 20 utterances of dysarthric female speakers. We propose a single stream branching of signal denoising through Sepformer and utilization of Swin-T for log-mel spectrogram classification. The method uses transformer-based engineering, showing promising results. While the EasyCall corpus offers more spoken utterances, future exploration will involve handling larger datasets and categories. Despite encouraging findings, there are numerous opportunities for further research and refinement.

In the next chapter, we have investigated objective 2, examining the effects of employing phonemes directly from raw audio, without the need for generating a text transcript. Additionally, we have explored the incorporation of phonemes and the development of a fusion framework for recognizing spoken words based on phonetic elements.

CHAPTER 4

DESIGN AND DEVELOPMENT OF FUSION FRAMEWORK FOR PHONEME-BASED SPOKEN WORD RECOGNITION FROM RAW AUDIO

Accented speech can pose challenges for ASR systems. Dialects, which are variations in language characterized by differences in words, phrasing, and influence by social and geographic factors, further complicate the task (Hinsvark et al. 2021: Holmes and Wilson (2017); Oh et al. 2021). Research indicates that DeepSpeech2-RNN holds the most accent-related information among ASR layers, especially in the initial stages (Prasad and Jyothi, 2020). Recent work by Winata et al. (2019) introduced model agnostic meta-learning (MAML) to accented ASR, demonstrating the potential for models to adapt to unfamiliar accents. Meta-learning techniques have also shown promise in training models to handle new accents with minimal exposure. While several methods like optimizers and various representations remain unexplored, accent embeddings, derived from late hidden layers in accent identification neural networks, have gained popularity for enhancing input features. Both hybrid acoustic models and E2E models have benefited from these embeddings (Jan et al. 2018; Rao et al. 2020).[2]

[2] The content of this chapter is published in:
"Early Fusion of Phone Embeddings for Recognition of Low-Resourced Accented Speech". - 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), https://doi.org/10.1109/AIST55798.2022.10064735.

The scarcity of adequate training data for accented speech has shown to result in higher WERs in ASR experiments (Lüdeling & Kytö, 2008; Lee (2007)). This chapter addresses the fundamental challenge associated with accented speech datasets: the lack of sufficient resources for model training.
The key highlights of this work are as follows:

- This chapter introduces a supervised technique for accented speech recognition, particularly in situations where resources are limited.
- Existing research has not extensively explored the use of phonology for understanding spoken text in accented speech recognition.
- The proposed approach employs early fusion of phone embeddings as a novel method for recognizing accented speech, especially from a small sample dataset.
- To ensure consistency, the vectors are concatenated and padded.
- The study aims to demonstrate the significant role that audio phonemes can play in accented speech recognition, even when training samples are limited in number.

The rest of the chapter is organized as follows. In Section 4.1, we outline the proposed dense architecture for the early fusion of phonemes. Section 4.2 details the experimental design and provides a comprehensive summary of the obtained results. The results are concluded in Section 4.3 outlines future research opportunities in the field.

4.1 MODEL DESCRIPTION

Within this section, we present our novel approach for early phoneme fusion in the context of accented speech recognition. Numerous researchers have embraced early fusion as a means to establish a cohesive representation by amalgamating multiple modalities (Zhao et al. 2018). This fusion of features is realized through concatenation and padding processes. Subsequently, these amalgamated features are input into a three-layered dense model, thus enhancing the performance of the accented speech recognition system. A visual representation of this innovative methodology can be observed in the accompanying Figure 4.1.

The initial step involves the transformation of raw audio into phonemes, a process executed through PocketSphinx (Mehra & Susan, 2021; Bojanowski et al. 2017; Gao et al. 2018; Huggins-Danies et al. 2006). Subsequently, these phonemes are converted into 300-dimensional vectors employing the FastText subword modeling technique. For this purpose, a dataset comprising two million word vectors from Web Crawl, generated using subword data (600 billion tokens), is employed. These word vectors are pre-trained representations and have been derived from extensive text corpora, including sources such as news collections, Wikipedia, and web crawls. Such pre-trained word representations are widely utilized in various text-based applications (Mikolov et al. 2017). To convert phonemes into vectors, subword modeling is employed. Phonetic characteristics extracted from the .wav raw audio files are stored as matrices, each having dimensions of 300 × 1, with the dimension measured per phone. To ensure uniformity, early fusion techniques are applied, involving the concatenation and padding of these vectors. This process results in vectors with dimensions of 43500 × 1. Subsequently, these uniform vectors are input into the 3-layered dense model, as depicted in Figure 4.2.

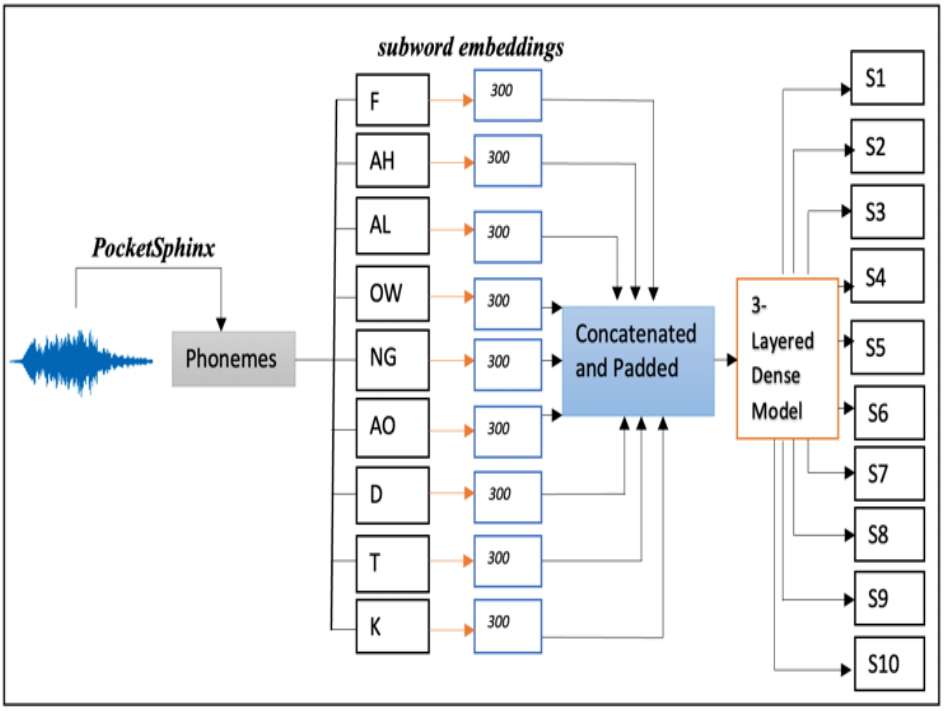


Figure 4.1. The block diagram of the proposed method

In Figure 4.1, the process of generating concatenated and padded features, resulting in a dimensionality of 43500 × 1, is illustrated. A sequence of phonemes with a vector dimensionality of 300 is concatenated as [phoneme^{1}, phoneme^{2}, …. , phoneme^{145}], where 145 represents the maximum length of the padded sequence of phonemes in a sentence. This leads to a vector of dimension 43500 × 1. The resulting feature vectors serve as input to the flatten layer, the first dense layer (512 units), the second dense layer (256 units), and the third dense layer (64 units) within our 3-layered dense model, as depicted in Figure 4.1. The model's output consists of probabilistic values, with the class label corresponding to highest probability (Esfe et al. 2022). For a comprehensive overview of the neural network architecture's parameters, please refer to Table 4.1.

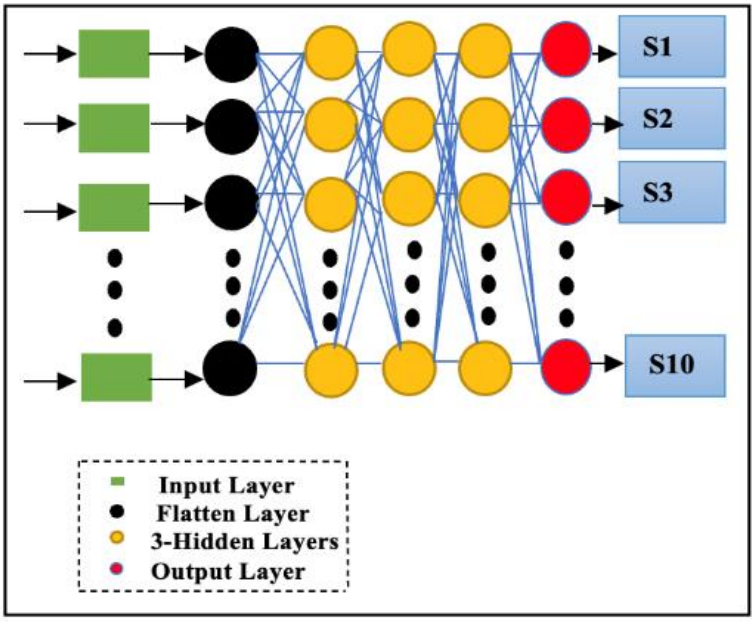


Figure 4.2. 3-layered dense model for classifying the fused phone embeddings

Table 4.1. Hyperparameters of our proposed approach

Framework	Values
Total Samples	240
Training Set Samples	168
Testing Samples	72
Dimensions	43500 × 1
Epochs	100
Activation function	ReLu
Dense layers	3

The 3-layered dense model, detailed in Table 4.1, takes the fused phone embeddings as its input. This model is responsible for training the padded features and comprises a flatten layer followed by three dense layers with unit sizes of (512, 256, and 64), all activated with the "ReLu" activation function. To regulate the learning rate in stochastic gradient descent, the Adam optimizer (Duchi et al. 2011) is utilized. This optimizer is renowned for its consistent performance across a range of tasks. The model employs a loss function known as sparse categorical cross-entropy.

4.2 EXPERIMENTAL RESULTS

4.2.1 DATASET

Our experimental data is drawn from the L2-ARCTIC English speech corpus designed for non-native speakers (Zhao et al. 2018). This dataset offers a balanced distribution in terms of gender and native languages (L1) and includes 26,867 utterances from 24 non-native speakers. The majority of the speakers from the CMU ARCTIC set recorded their entire speech events. Collectively, these speakers contributed an average of 67.7 minutes of speech each, resulting in a corpus duration of 27.1 hours, with a standard utterance length of 8.6 minutes. On average, each utterance spans 3.6 seconds, and the intervals between speech samples are typically less than 100 milliseconds. The dataset comprises approximately 238,702 word segments and 851,830 phone segments, with each speech sample containing an average of nine words (excluding silence segments). Annotators meticulously reviewed 3.599 utterances and identified 1,092 phone additions, 3,420 phone deletions, and 14,098 phone substitutions. Some sentences are truncated due to incomplete readings by certain speakers or subpar recording quality. For the task of spoken sentence categorization, we adapted the L2-ARCTIC dataset, which initially featured speakers from six different languages-Korean, Arabic, Hindi, Chinese, Spanish, and Vietnamese. Each language category included four distinct speakers with diverse accents. However, the dataset had primarily been designed for research purposes related to voice and accent conversion, as well as mispronunciation identification. Consequently, the number of available audio samples per sentence category was relatively limited. Due to the dataset's handling and management, it can be considered under-resourced. Specifically, each sentence category comprises only 24 speakers, resulting in a dataset size of 240 samples for ten sentence

classes. To conduct our experiments, we adopted a 70:30 split ratio, segregating the dataset into training and testing sets, and implemented a 3-fold cross-validation approach.

4.2.2 Experimental Setup

The experiments are conducted on a system running MAC OS High Sierra, equipped with an Intel Core i5 processor featuring Intel Graphics and operating at a clock speed of 1.8 GHz. The experiments are implemented using Python version 3.10.4. We adopted a 70:30 train-to-test split ratio, employing a 3-fold cross validation approach to ensure robust evaluation. For optimization, we utilized the Adam optimizer, with a batch size of 16, an extensive training duration of 100 epochs, and a learning rate set at 0.01. The non-linearity in the model is introduced via the "ReLu" activation function.

4.3 Results And Discussions

We employed PocketSphinx to extract phonemes from the raw audio data. After phoneme extraction, we proceeded to normalize the phonemes, which involved eliminating noise phones such as SPN and NSN. This pre-processing step effectively removed signal noise from the retrieved phonemes. To represent the phonemes as vectors, we employed subword embeddings, with each phoneme being transformed into a 300-dimensional vector. Additionally, padding was applied to ensure uniformity in the vector dimensions, resulting in padded vectors with dimensions of (43500×1). These padded vectors served as input for the deep dense layered model, as illustrated in Figure 4.1. The experiments are conducted with three cross-validations, yielding an average testing accuracy of 49.01% for the proposed method. Given the low-resource nature of the dataset and the presence of accented speech, the achieved accuracy is within the expected lower range.

Table 4.2 provides a comprehensive classification report, encompassing F1-scores and accuracies for the ten sentence classes evaluated using our proposed methodology. The results reveal variations in the ease or difficulty of classifying different sentences. Notably, our method achieved a remarkable 100% accuracy in recognizing sentence category -5, specifically for the sentence "Will we ever forget it". However, it struggled with sentence category -6, where the sentence was "God bless 'em, I hope I'll go on seeing them forever".

However, it struggle with sentence category -6, where the sentence was “God bless ‘em, I hope I’ll go on seeing them forever”. This performance discrepancy reflects the varying levels of success in recognizing sentences, with only a limited number of sentences being correctly identified. This challenge primarily arises from the complexity of the dataset, characterized by its challenging accents and distinctive speech patterns. In Figure 4.3, the confusion matrix for the early fusion model is presented. It’s evident from this visualization that our proposed technique achieves the highest accuracies for sentence category -5. It's worth noting that with a larger dataset, the model’s performance could have been significantly improved. The accuracy levels for accented speech from non-native speakers tend to be relatively lower due to the dataset’s limited size. However, these results still provide valuable insights into the effectiveness of different models in handling such challenging linguistic variations. The ROC plots in Figure 4.4 align closely with the class-wise F1-scores and accuracies presented in Table 4.2. In Table 4.3, various evaluation metrics for different sentence classes are displayed. These metrics include sensitivity, specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Positive Rate (FPR), False Negative Rate (FNR), and False Discovery Rate (FDR). It’s noteworthy that both specificity and NPV exhibit high values across all classes, while sensitivity and PPV are particularly high for specific classes, such as sentence categories -5, 4, and 8. FPR, FNR, and FDR values are consistently low, with FPR values exhibiting low rates across all classes. The most notable performance is observed for sentence category -5.

Table 4.2. F1-Score and accuracy per sentence category

Sentence	Classification Results			
	precision	*recall*	*F1-score*	*Accuracy*
S1	0.50	0.29	0.36	50%
S2	0.43	0.60	0.50	57%
S3	0.20	0.50	0.29	80%
S4	0.42	0.83	0.56	58%
S5	1.00	1.00	1.00	100%
S6	0.00	0.00	0.00	0%
S7	0.67	0.40	0.50	33%
S8	0.67	0.75	0.71	33%
S9	0.40	0.25	0.31	60%
S10	0.75	0.33	0.46	25%

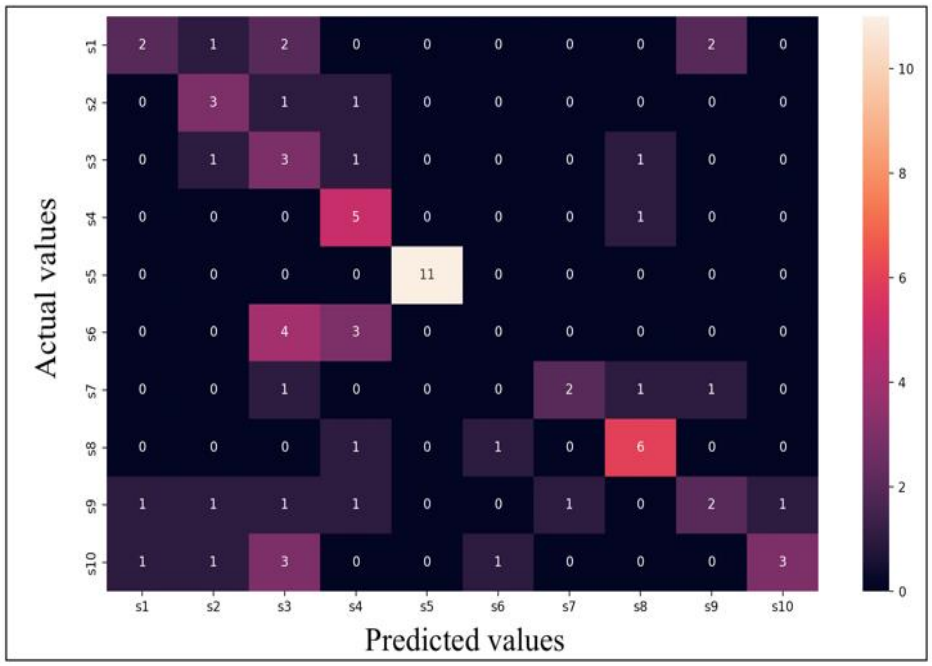


Figure 4.3. Confusion matrix of proposed method

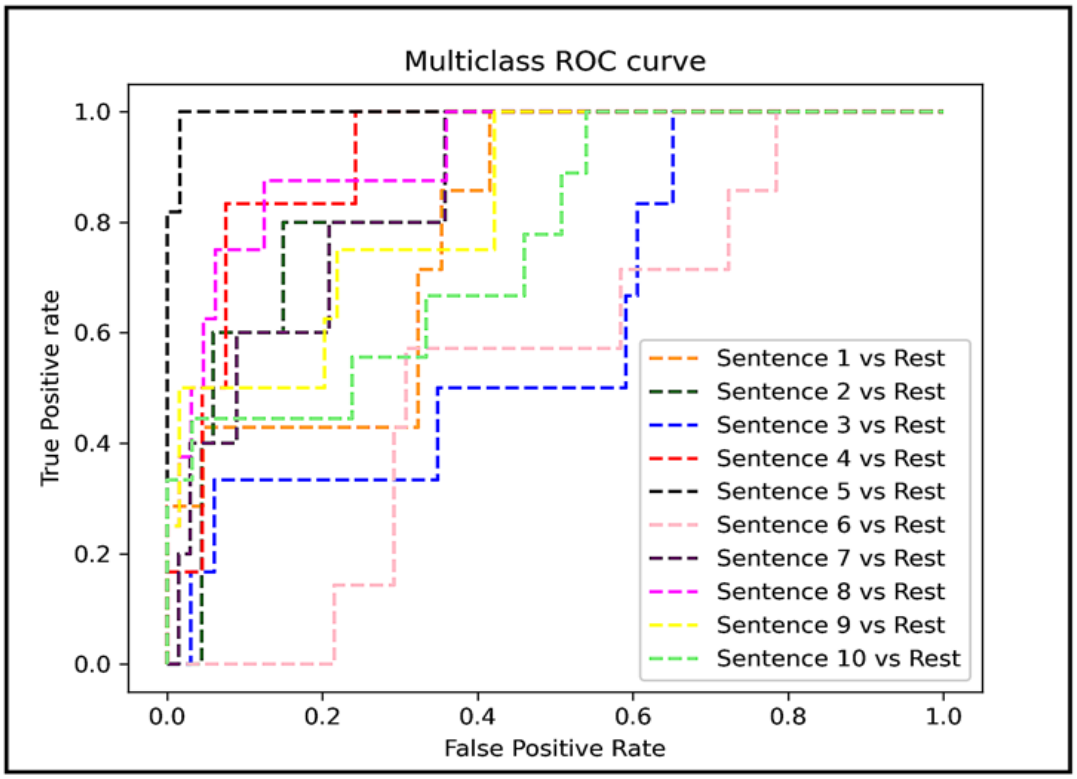


Figure 4.4. Multiclass classification evaluation with receiver operating characteristic curve

We conducted a comparative analysis of our proposed method against other approaches, including a CNN using MFCC features as 2D input. Notably, our method demonstrated a

superior performance in accented speech recognition, outperforming the MFCC with CNN by an impressive margin of 4.18%. Additionally, we benchmarked our approach against a technique employing 300-dimensional text-glove word embeddings, which are classified using a LSTM model. With a configuration of 100 epochs, 128 neurons, the Adam optimizer, and the tanh activation function, our method achieved an accuracy of 42.99% on the accented dataset, surpassing the LSTM-text-glove approach by a substantial 6.02%. It's worth highlighting that while CNN has shown effectiveness in sound classification when using Mel Spectrogram as an input feature, our proposed approach outperformed Mel Spectrogram with LSTM by a notable margin of 5.3%. For detailed performance comparison, please refer to Table 4.3. Our results undeniably establish the superiority of our method in the context of limited resource availability. In situations where classifying spoken sentences based on raw audio is challenged by a scarcity of training and testing examples, achieving satisfactory results can be a formidable task. However, our proposed approach shines through these challenges, consistently delivering the best performance. Looking ahead, the future of our work is poised for further exploration and expansion. The focal point is on investigating a broader range of fusion models tailored for low-resourced accented speech corpora.

Table 4.3. Classification report per sentence category

Sentences	Classification Results						
	Sensitivity	*Specificity*	*PPV*	*NPV*	*FPR*	*FNR*	*FDR*
S1	0.28	0.97	0.50	0.93	0.30	0.71	0.50
S2	0.60	0.94	0.42	0.97	0.05	0.40	0.58
S3	0.50	0.82	0.20	0.95	0.19	0.50	0.80
S4	0.83	0.90	0.42	0.98	0.11	0.17	0.58
S5	1.00	1.00	1.00	1.00	0.00	0.00	0.00
S6	0.00	0.97	0.00	0.90	0.03	1.00	1.00
S7	0.40	0.98	0.66	0.96	0.02	0.60	0.33
S8	0.75	0.95	0.66	0.96	0.05	0.25	0.33
S9	0.25	0.95	0.40	0.91	0.05	0.75	0.60
S10	0.33	0.98	0.75	0.91	0.02	0.67	0.25

4.4 SIGNIFICANT OUTCOMES

This study delves into the classification of low-resourced accented speech through the innovative application of early fusion of phoneme embeddings and a dense model. Notably, our approach is pioneering in that it introduces the combination of early phone fusion with

accented speech, a novel endeavour within the context of low-resourced datasets. The amalgamation of early fusion of phoneme embeddings and a 3-layered dense model significantly augments the performance of low-resourced accented speech recognition. We conducted our experiments using the L2-ARCTIC accented speech dataset, featuring audio recordings from 24 non-native English speakers and encompassing 10 sentence categories. Our proposed methodology has yielded commendable results in the classification of spoken sentences. Our future work will focus on further exploring diverse fusion models tailored to low-resourced accented speech corpora, thereby extending the scope of this research. In the next chapter, we have explored objective 3, investigating the impact of using phonemes and morphemes extracted from raw audio speech transcriptions. The focus is on assessing the importance of a fusion framework for recognizing spoken words based on these phonetic and morphological element.

CHAPTER 5

DESIGN AND DEVELOPMENT OF FUSION FRAMEWORK FOR PHONEME- AND MORPHEME-BASED SPOKEN WORD RECOGNITION FROM SPEECH TRANSCRIPTIONS

In this thesis, we present an innovative unsupervised technique designed to rectify severely flawed speech transcriptions. This approach employs a decision-level fusion strategy, combining stemming and a two-way phoneme pruning process. This combination of techniques seeks to enhance the accuracy and quality of transcriptions in the context of challenging or imperfect speech data. This chapter has been dedicated to the enhancement of speech transcription through the implementation of decision fusion, a notably effective approach. We conducted experiments encompassing 500 diverse word categories, involving multiple speakers representing both male and female genders. Although the transcriptions extracted from Google API may not achieved exceptionally high accuracy, they have demonstrated practical utility, particularly in real-time applications. The key highlights are:

- Introduction of an unsupervised approach for improving highly imperfect speech transcriptions through decision-level fusion of stemming and two-way phoneme pruning.
- Utilization of the Ffmpeg framework for extracting audio from videos, followed by audio-to-text transcription using Google API. Utilization of the LRW dataset, featuring 500 word categories and 50 videos per class in mp4 format, with each video comprising 29 frames.[3]

[3] The content of this chapter is published in:
"Improving word recognition in speech transcriptions by decision-level fusion of stemming and two-way phoneme pruning. - Advanced Computing: 10th International Conference" IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10, https://doi.org/10.1007/978-981-16-0401-0_19.

- Aiming to enhance baseline accuracy, resulting in a notable improvement from 9.34%.
- Incorporation of stemming, phoneme extraction, filtering, and pruning techniques in the approach, achieving 23.34% accuracy in word recognition.
- Leveraging the CMU pronouncing dictionary for G2P conversion.
- Introduction of a two-way phoneme pruning method, involving non-sequential steps for filtering and pruning phonemes containing vowels, plosives, and fricatives.
- Application of decision-level fusion, ultimately elevating word recognition rates up to 32.96%.

This chapter introduces a comprehensive approach that combines stemming techniques with two distinct phases of phoneme filtering and pruning to facilitate word recognition and error correction in flawed speech transcriptions. The chapter's structure is as follows: Section 5.1 outlines the proposed methodology, Section 5.2 provides an in-depth analysis of the experimental outcomes, and Section 5.3 presents the significant outcome remarks.

5.1 PROPOSED APPROACH FOR CORRECTING IMPERFECT SPEECH TRANSCRIPTION

5.1.1 TEXT PRE-PROCESSING

The text transcript is relatively concise, exemplified by the word category "significant" in Figure 5.1. Speech transcript normalization is a crucial step involving data cleansing, which entails the removal of undesirable elements such as stop words and punctuation. Additionally, it encompasses the conversion of numeral values into their textual representations and transforming all words into lowercase for improved readability. Further, the process includes sentence tokenization to facilitate proper content comprehension by distinguishing individual words. Text filtering plays a pivotal role in expediting processing while concurrently reducing the document's overall size. The elimination of stop words is a crucial step since these words, such as "a", "the", "an", "of", "like", and "for", hold minimal significance in the context of information retrieval. Furthermore, we've undertaken the task of rephrasing commonly used expressions, such as "couldn't", by presenting them in their grammatical forms, like "could not". Tokens that incorporate symbols like ".", "!", "#", and "$" are tailored to fit the content's requirements and, where necessary, are either converted to word form or removed altogether, as illustrated in Figure 5.2.

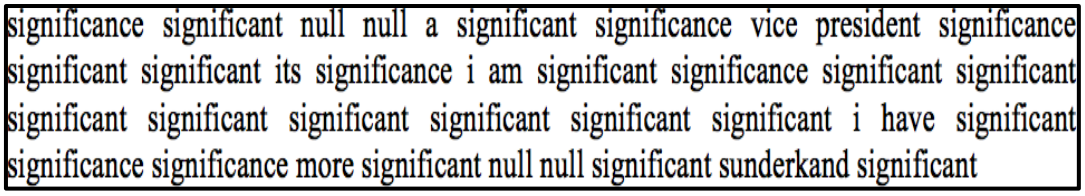


Figure 5.1. Text transcription after pre-processing

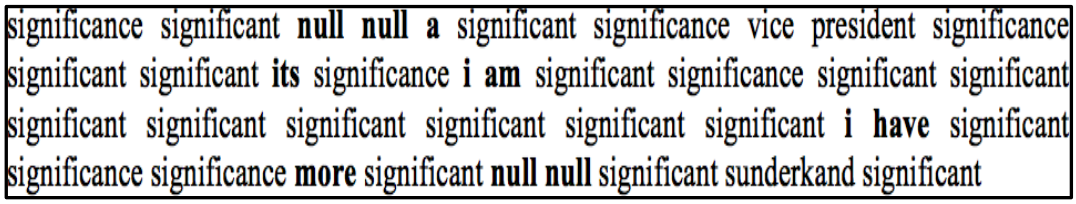


Figure 5.2. After text normalization (removing all these bold letter words) from a transcription

5.1.2 STEMMING

Stemming, employed in information retrieval and linguistic morphology, serves as a technique to truncate word suffixes and reduce them to their fundamental root form, also referred to as the base form. This process finds wide application in text and NLP. For instance, if a word concludes with “ed”, “ies”, “ing”, or “ly”, the trailing portion of the word is removed to derive its root or base form. One of the tools used for stemming is the Porter Stemmer, which, as illustrated in Figure 5.3, may not always align with the morphological root of words. This stemmer algorithm delineates the process of eliminating inflectional endings and common morphs from words, aiding in text normalization within information retrieval systems (Porter, 1980). Notably, the Porter stemmer exhibits a milder approach compared to the Lancaster stemmer, which tends to trim more of the valid text. In the context of linguistic morphology, stemming revolves around the quest for a word’s root or base. Conversely, lemmatization focuses on identifying the lemma from a lexicon containing words with identical senses. The base or word-form can be inferred or inflected.

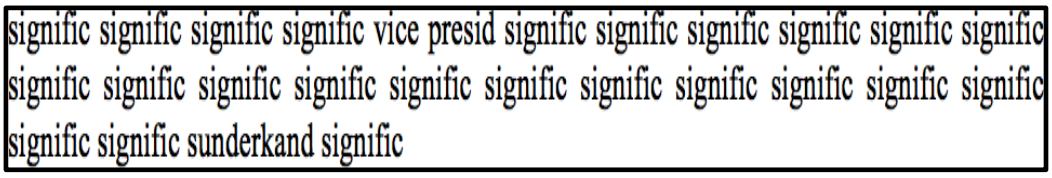


Figure 5.3. After applying stemming using Porter stemmer

In our LRW dataset, the keyword "absolutely" possesses a lemma as "absolutely" and a stem as "absolut". As we examine the text transcriptions, the prevalence of the stem "absolut" surpasses that of the lemma "absolutely". Therefore, the stemmer brings the majority of words closer to their respective categories. This proximity results in higher classification scores compared to experiments that adhere to matching words in their original form.

5.1.3 PHONEME EXTRACTION FROM TEXT

Phonemes, within the realm of linguistics, constitute the smallest speech units that distinguish sounds or groups of sounds, impacting both pronunciation and meaning. The way a phoneme is pronounced may vary depending on the surrounding letters and their influence on its representation (Alsharhan & Ramsay, 2019). These phonemes encapsulate the sound of speech, encompassing aspects like stress, articulation, and intonation. In our work, we harnessed the CMU pronunciation dictionary, an extensive lexicon boasting over 125,000 words and their corresponding phonetic transcriptions.

Following the normalization of text, we proceeded to extract a collection of phonemes from the aligned text transcript. This collection focuses solely on the standard sound patterns and stress configurations observed in syllables, words, and phrases. For instance, the phonetic transcription of the word "about" appears as "AH0 B AW1 T", wherein "0" signifies the absence of stress, "1" denotes primary stress, and "2" indicates secondary stress. We meticulously filtered out all numerical values. Consequently, the text underwent segmentation into words and was further represented by its constituent phonemes, as exemplified in Figure 5.4.

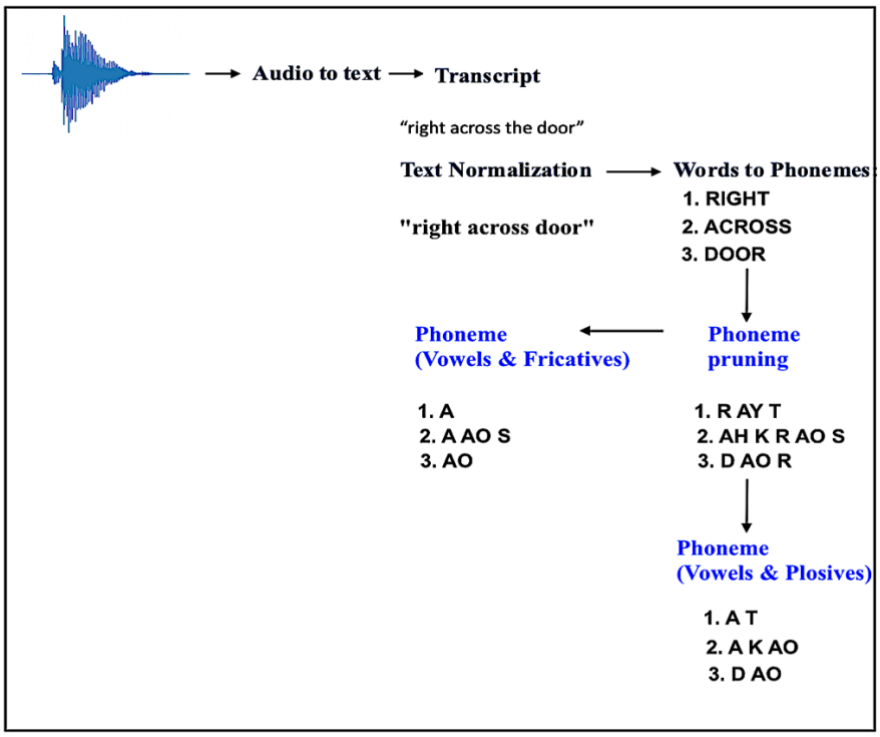


Figure 5.4. Overview of phoneme filtering and pruning from a sample phrase. Text phoneme is filtered and pruned to include plosives and vowels and alternatively, to include vowels and fricatives in the same manner

Distinct speakers bring forth varying pronunciations, making the quest for a precise transcript match a formidable challenge (Hazen, 2006). The CMU pronouncing dictionary adheres to American English standards when approximating text pronunciations. Notably, this dictionary aligns with the standardized representation found in the IPA, a system that captures the sounds of spoken language across the globe.

Phoneme filtering is an essential step involving the categorization of phonemes based on their content, specifically whether they contain vowels, plosives, or fricatives. This process not only reduces the dataset's size but also enhances category identification, as elucidated by the sequence of phonemes depicted in Figure 5.5. These phonemes are extracted from the normalized text presented in Figure 5.2. Plosives, also known as stop or oral consonants, operate by obstructing the vocal tract, momentarily halting the airflow. They encompass both voiced and voiceless consonants. Voiced plosives include "b", "d", and "g", while voiceless or unvoiced plosives consist of "p", "t", "k'. They are also referred to as glottal stops. Fricatives, which primarily comprise voiced consonants characterized by high energy and

amplitude, encompass examples such as “f”, “s”, “v”, and “z”. Due to their robust amplitude, they are readily detectable. The IPA has introduced a phonetic chart designed to distinguish the sounds of various human languages. Consonants, produced within the vocal tract, notably the oral tract (encompassing the mouth and pharynx), play a pivotal role in constructing spoken language. Consonants can be further categorized as labials, coronals, radicals, and dorsals. Vowels contribute to high pitch and amplitude, often rendering them correctly detectable. Based on the manner of speech production, sounds are classified as nasal if originating from the nose, while sounds created by obstructed airflow are referred to as stops, plosives, or orals. Fricatives, typified by consonants like “f”, “s”, “v”, and “z”, involve the placement of the lower lip against the upper teeth during their production. Trills are akin to taps, while flaps are sounds generated through the interaction of active and passive articulators (Richard and Makhoul, 1975.

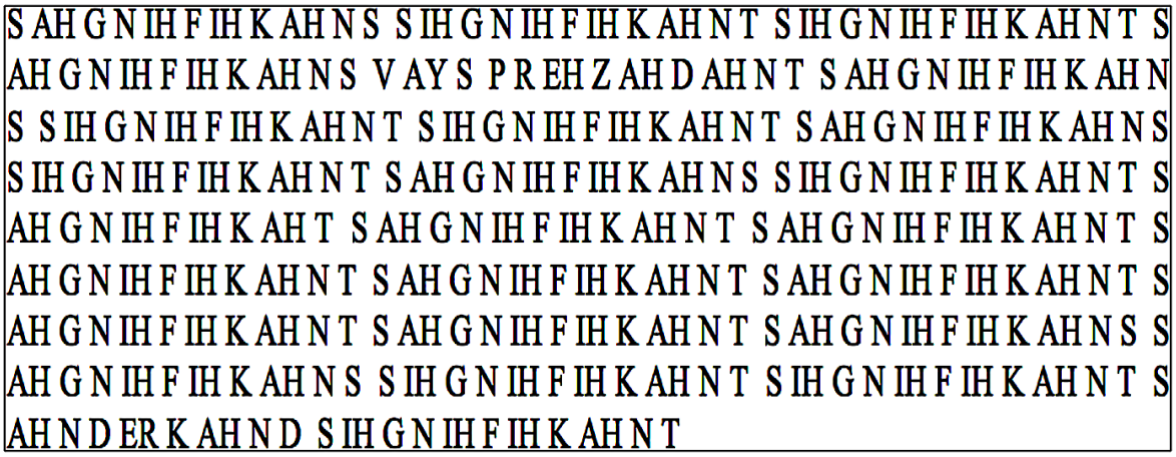


Figure 5.5. After phoneme extraction and filtering

5.1.4 Decision Fusion Of Stemming And Phoneme

Following the compilation of outcomes from pure stemming and phoneme filtering, we proceeded with a phoneme pruning process. This pruning involved two distinct stages, where we retained phonemes containing either (i) vowels or plosives or (ii) vowels or fricatives. These two stages were executed in a non-sequential manner, constituting what we refer to as the two-way phoneme pruning scheme.

Stage I: Phoneme pruning using Vowels and Plosives
Stage II: Phoneme pruning using Vowels and Fricatives

To harness the collective strengths of all these elements, we amalgamated the results obtained from stemming, Stage I, and Stage II of the two-way phoneme pruning through a decision-level score fusion. Under this scheme, if any of the stages, namely stemming, or the presence of Vowels + Plosives, or Vowels + Fricatives, successfully identifies a particular word in the transcript, we consider the word as identified. This approach ensures comprehensive recognition.

5.2 EXPERIMENTAL RESULTS

5.2.1 DATASETS

The Lip Reading in the Wild (LRW) dataset (Yang et al. 2019) poses a significant challenge for speech recognition in the real-world conditions and has served as a catalyst for numerous studies in audio-visual speech recognition (Haubold & Kender, 2007; Torfi et a., 2017). In our research, we focused on extracting and processing the audio track from this dataset to generate speech transcriptions. The LRW dataset is comprised of 500 distinct word categories, each containing 50 samples. For our unsupervised experiments, we exclusively utilized the testing data. These videos are all in .MP4 format and consist of 29 frames, each lasting 1.16 seconds, with the word typically appearing in the middle of the video. The specific word lengths and details are documented in the metadata. To extract audio from the videos, we employed the Ffmpeg framework, a rapid and versatile multimedia file converter that ensures no loss in quality during format conversion.

5.2.2 RESULTS OF THE PROPOSED APPROACH USING DECISION-LEVEL FUSION

The experiments are conducted using Python 3.7.4 on a Mac OS High Sierra system equipped with an Intel Core i5 processor and Intel HD Graphics 6000 (1536 MB) with a clock speed of 1.8 GHz. The processing time for a single audio file was approximately 29 seconds. In our proposed methodology, we engaged in phoneme filtering and pruning, selecting vowels and plosives in Stage 1 and vowels and fricatives in Stage 2. After implementing stemming, we achieved a word recognition rate of 23.34%. Stage 1 of phoneme pruning resulted in a recognition rate of 27.67%, while Stage 2 yielded 28.23%.

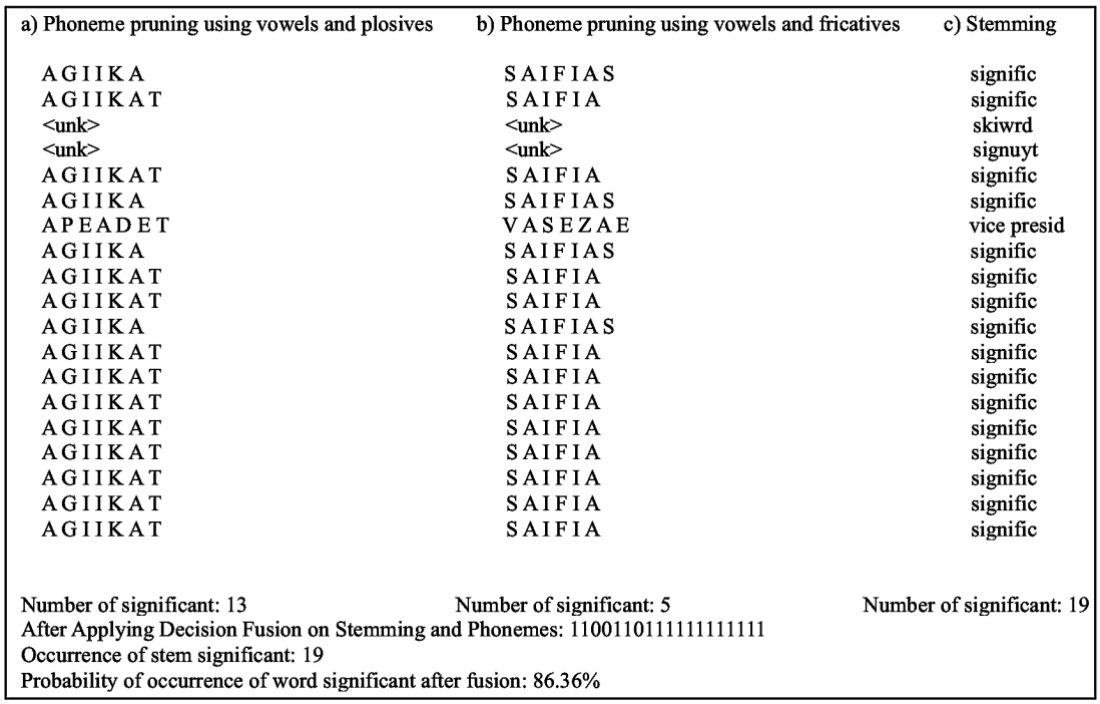


Figure 5.6. Decision fusion of stemming and two-way phoneme pruning

Notably, when we applied decision-level fusion to combine stemming, Stage 1, and Stage 2 of phoneme pruning, the word recognition rate significantly improved to 32.96%. This represents a substantial enhancement over the baseline method, which yielded a recognition rate of 9.36%, as elaborated in the subsequent subsection. A demonstration of decision-level fusion for the word category "significant" can be observed in Figure 5.6. A concise summary of all results is provided in Table 5.1, accompanied by relevant references to the literature that highlights the utilization of individual components in speech text understanding. Additionally, we conducted a comparison of our approach with two established automated spelling correction tools, namely, autocorrect (Rayson et al. 1998) and Symspell (Gupta, 2006), the code for which is accessible online at respectively.

Table 5.1 Comparison of various methodologies on LRW dataset

Methods	Accuracy
Baseline (Bahl et al. 1984)	9.34%
Stemming [(Mathew et al. 2016)	23.34%
Phoneme pruning (Vowels & Plosives) (Hemakumar, 2011)	27.67%
Phoneme pruning (Vowels & Fricatives) (Haubold and Kender, 2017)	28.23%
Autocorrect (Rayson, 1998)	21.50%
Symspell (Gupta, 2020)	25.16%
Decision Fusion of Stemming and two-way Phoneme pruning	32.96%

5.2.3 SLIDING TEXT WINDOW- THE BASELINE APPROACH

In the baseline method, we employ a window that traverses the text, breaking it into tokens, with each token corresponding to a word within the sentence. Upon identifying the category word in a sentence, the text window is free to move either to the next sentence or to the subsequent line in the text document. As we navigate through the sentence, word by word, and traverse through lines, we calculate the frequency of occurrences of the category word, as illustrated in Figure 5.7. This approach effectively mitigates issues related to the duplication and redundancy of the same category word within a sentence.

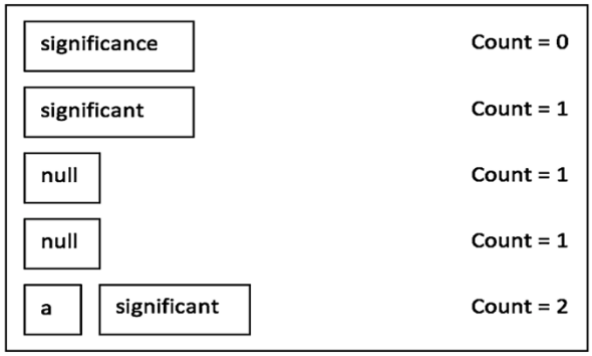


Figure 5.7. Sliding text window to search for the keyword “significant” in-text transcription

5.2.4 TEST CASES

Let’s examine a few test cases to gain a more precise understanding of the role of decision fusion:

Test Case 1: In certain cases, the stem alone is adequate for word recognition. For example, in the category “announced”, we have the following probabilities: stem (root form “announce”) – 6%, vowels + plosives -4%, vowels + fricatives -2%. Decision fusion in this case strongly favours the stem, as removing “ed” from “announced” increases the stem’s probability by 4%. For vowels + plosives, the CMU dictionary pronunciation is “AH”, “N”, “AW”, “N”, “S”, and “T”, with the desired pattern “AAT” in the transcript, leading to a 2% probability increase. In the case of vowels + fricatives, the CMU dictionary pronunciation is “AH”, “N”, “AW”, “N”, “S”, “T”, with the desired pattern “AAWS” in the transcript,

resulting in the same probability as the baseline. The stem outperforms the other two cases in this specific test case.

Test Case 2: Let's consider another category, "agreement", where the probabilities are as follows: stem -44%, vowels + plosives -88%, vowels + fricatives -92%. Here, decision fusion strongly favours vowels + fricatives. For the stem, the probability remains the same as the baseline. However, for vowels + plosives, the CMU dictionary pronunciation is "AH", "G", "R", "IY", "M", "AH", "N", "T", with the desired pattern "AGIAT" in the transcript, leading to a 22% probability increase. In the case of vowels + fricatives, the CMU dictionary pronunciation is "AH", "G", "R", "IY", "M", "AH", "N", "T", with the desired pattern "AI" in the transcript, resulting in a 70% probability increase. Vowels + fricatives perform best in this test case.

Test Case 3: In the category "affairs", the probabilities are as follows: stem -4%, vowels + plosives -20%, vowels + fricatives -6%. Decision fusion strongly favours vowels + plosives. After removing "s" from "affairs", the probability of the stem "affair" remains the same as the baseline. For vowels + plosives, the CMU dictionary pronunciation is "AH", "F", "EH", "R", "Z", with the desired pattern "AE" in the transcript, leading to a 16% probability increase. In the case of vowels + fricatives, the CMU dictionary pronunciation is "AH", "F", "EH", "R", "Z", with the desired pattern "AFERZ" in the transcript, resulting in a 2% probability increase. Vowels + plosives yield the best results in this particular test case.

These test cases demonstrate how decision fusion effectively combines the strengths of various methods to provide the highest accuracy for word recognition across different categories.

5.3 Significant Outcomes

Our approach involves the fusion of two effective techniques, stemming and two-way phoneme pruning, for improving word recognition accuracy in highly imperfect speech transcriptions extracted from the LRW dataset in mp4 format. We initiate the process by extracting audio samples from the videos using the Ffmpeg framework. Subsequently, we convert the audio speech into text transcriptions using the publicly available Google API, which has versatile applications in speech adaption, speech transcription, and real-time speech recognition. To evaluate our results, we begin with a baseline comparison, which

involves simple string matching to identify word categories in the text transcription. Our first step is text normalization and speech adaption, which entails the removal of stop words, the most frequent and extraneous words in the text, to expedite text processing. Next, we apply stemming to derive the root form of each word and compare it against various word categories. Concurrently, we convert each word into phonemes using the CMU pronouncing dictionary. We then map the text transcript to phonemes and proceed to apply phoneme filtering, where we selectively filter out phonemes containing vowels, plosives, or fricatives.

The phoneme pruning process comprises two non-sequential stages: Stage I involves phoneme pruning using vowels and plosives, while stage II focuses on phoneme pruning using vowels and fricatives. Subsequently, we accumulate results from these three methods and apply decision fusion to ascertain whether any of these methods successfully detect the occurrence of the word. The proposed fusion method proves to be highly effective, surpassing existing SOTA techniques. As a result, word recognition accuracy is significantly enhanced, elevating it from a baseline accuracy of 9.34% to an impressive 32.96% using our fusion approach.

In the next chapter, we have delved into objective 4, exploring the effects of employing a classification framework for phonological and morphological features using pre-trained networks in the realm of SWR.

CHAPTER 6

DESIGN AND DEVELOPMENT OF CLASSIFICATION FRAMEWORK FOR PHONOLOGICAL- AND MORPHOLOGICAL FEATURES USING PRE-TRAINED NETWORKS FOR SPOKEN WORD RECOGNITION

In this thesis, we proposed two different novel techniques. First, we introduce a method to recognize spoken words using minimal input data. Previous research has often overlooked the potential of using linguistic elements like morphemes and phonemes to understand spoken text. To address this, we present a late fusion approach that combines phone embeddings and bigram embeddings for SWR. We work with audio samples in .OPUS format and extract text transcripts from them using a pre-trained English classifier called xlsr-Wav2Vec2-53. From these transcripts, we obtain phonemes using the CMU pronouncing dictionary. These phonemes are then converted into vectors using ML language-agnostic sentence embeddings. We also extract bigrams from the text transcripts and vectorize them in a similar manner. Both the phoneme embeddings and morpheme embeddings are fed into a 5-layered dense batch normalization model. These results outperform existing techniques. Our work demonstrates that linguistic elements, such as phonemes and morphemes, can significantly contribute to SWR, especially when training data is limited and imbalanced. We also show that using text-transcription features from pre-trained models can be more successful than traditional audio-based feature modeling.

The contents of this chapter are submitted/accepted/under review in:
"Spoken Word Recognition for Asian Languages using Transformers" - Computer Speech and Language. (IF: 4.3).
&
"Speaker Independent Recognition of Low-Resourced Multilingual Arabic Spoken Words through Hybrid Fusion" – accepted in Multimedia Tools and Applications, https://doi.org/10.1007/s11042-024-18804-w. (IF: 3.6).

The main contributions of this chapter are outlined as follows:

- The ML landscape of spoken word classification presents a unique challenge in the area of NLP.
- We propose a supervised approach to tackle this challenge by leveraging the power of transformer models for the extraction of text transcripts from audio data in Arabic, Tamil, and Vietnamese.
- In our pursuit of advancing the SOTA, we present a novel late fusion technique that synergistically combines phone embeddings and bigrams.
- Our approach leverages both phone embeddings and bigrams, capitalizing on their individual strengths to enhance SWR. By introducing a late fusion technique, we enable these two diverse features to complement each other, resulting in a more comprehensive representation of spoken words. The late fusion model elegantly integrates these features, paving the way for enhanced accuracy in the recognition process.
- To process both phone embeddings and morph embeddings, we employ a 5-layered dense batch normalization model. This architecture is meticulously designed to extract intricate patterns and relationships within the input data. By utilizing batch normalization at each layer, we ensure model stability and faster convergence. The model serves as the backbone for our late fusion approach, allowing it to thrive in the complex area of ML SWR.
- Our efforts culminate in superior accuracy within the ML spoken dataset, surpassing existing SOTA methods. We benchmark our technique against a set of ten distinct word categories, each representing unique linguistic challenges. Through extensive experimentation and fine-tuning, we consistently outperform previous recognition models, underscoring the effectiveness and versatility of our late fusion approach.
- This innovation is not confined to a specific language or dialect; it has the potential to revolutionize SWR across diverse languages. Our technique is a testament ot the evolving landscape of DL in speech processing, setting a new standard for accuracy and adaptability.
- In summary, our late fusion technique, in combination with a robust DL architecture, propels the field of ML SWR to new heights. We anticipate that this approach drive

advancements in speech technology and foster a deeper understanding of spoken language across the world's linguistic diversity.

The structure of the chapter is as follows: Section 6.1 presents the proposed dense architecture for late fusion of phonemes and bigrams. Section 6.2 outlines the experimental design and summarizes the results. Section 6.3 concludes the chapter and suggests potential directions for future research.

6.1 PROPOSED APPROACH

In this section, we present our late fusion method designed for the recognition of ML spoken words. Table 6.1 offers essential details regarding the training-to-testing sample ratio within the ML spoken words corpus.

Table 6.1 Per category train test samples

Languages	Train – Test		
	Arabic	*Tamil*	*Vietnamese*
category 1	4-2	3-2	4-3
category 2	13-6	3-2	4-2
category 3	3-2	2-2	4-3
category 4	4-2	3-2	3-2
category 5	8-4	3-2	4-2
category 6	4-2	46-20	4-2
category 7	7-3	7-3	10-5
category 8	9-5	11-6	7-3
category 9	5-3	4-2	9-4
category 10	5-3	4-2	6-3

We have employed an under-resourced and significantly imbalanced dataset encompassing Arabic, Tamil, and Vietnamese languages. As evident from the information in Table 6.1, it is apparent that we are working with a low-resourced dataset. Algorithm provides a detailed explanation of our E2E approach.

Algorithm: Linguistic feature extraction, classification with deep neural network, and Fusion

Input: Train Set X_{train}, Test Set X_{test}, Train Labels Y_{train}, Test Labels Y_{test}
Output: ACCURACY

1. **For each** audio file **in** [X_{train}, X_{test}] **do**
   TEXT_TRANSCRIPT ← Extract Text-Transcript (using English Large xlsr-Wav2Vec2-53 pre-trained classifier)
   End for
2. Convert TEXT_TRANSCRIPT to phonemes and bigrams
3. **For each** Phonemes | Bigrams **in** [X_{train}, X_{test}] **do**
   LINGUISTIC_FEATURE ← Vectorize Phonemes | Bigrams (using language agnostics BERT sentence embedding (LaBSE) model)
   End for
4. Instantiate two separate deep neural networks for training on samples of LINGUISTIC_FEATURE, belonging to X_{train} having labels $Y_{train.}$ Use deep neural networks to predict the category label of the X_{test} for the two characteristics.
5. Create lists to store the posterior category probabilities that the two models have calculated.
6. SPOKEN_WORD_CATEGORY_PROBABILITY ← Fusion of posterior probabilities for each spoken word category
7. Y_{pred}← argmax (SPOKEN_WORD_CATEGORY_PROBABILITY)
8. ACCURACY ← Calculate accuracy using (Y_{pred}, Y_{test})

6.1.1 ENGLISH LARGE XLSR-WAV2VEC2-53

We selected the Facebook/wav2vec2-large-xlsr-53 model, which is based on the Common Voice 6.1 language corpus, as our primary language model. This model has achieved a SOTA WER of 14.01% on the Common Voice dataset, making it the ideal choice for English ASR. In our proposed method, we employ the English Facebook/wav2vec2-large-xlsr-53 model to extract text transcripts from the raw audio. For each language in consideration, we extract the respective text transcript. These text transcripts are then processed in two distinct ways. First they are converted into phonemes using grapheme-2-phoneme modeling, which is elaborated on in the following section. Concurrently, the extracted text transcripts are also transformed into bigrams using n-grams NLP.

6.1.2 GRAPHEME-TO-PHONEME MODELING

This module serves the purpose of translating English spelling graphemes into phonemes, which are representations of word pronunciations. This capability is of significant

importance in various applications, notably in the field of voice synthesis, as highlighted by He and Deng (2022). In contrast to many other languages, like Spanish or German, where the pronunciation of a word can often be reliably deduced from its spelling, English words tend to exhibit frequent deviations from expected pronunciation, as noted by Bisani and Ney (2008). Consequently, to determine the correct pronunciation of an English word, it is highly advisable to consult a resource like the CMU pronouncing dictionary. In our method, the text transcript is subjected to a conversion process that maps graphemes to phonemes based on the CMU pronouncing dictionary. Subsequently, these phonemes are further transformed into embeddings using LaBSE, a procedure that is detailed in the following section. This approach ensures that we accurately capture the phonetic representations of the words, enabling more precise and effective speech recognition.

6.1.3 LANGUAGE AGNOSTICS BERT SENTENCE EMBEDDINGS (LABSE)

We harnessed the power of LaBSE to convert the extracted phonemes and morphemes into vector representations. LaBSE, a model introduced by Feng et al. in 2020, boasts support for a remarkable 109 languages. This ML embedding model is an invaluable tool, seamlessly blending semantic information to facilitate language understanding and offering the capability to encode text from diverse languages into a shared embedding space. This flexibility makes LaBSE suitable for a wide array of downstream tasks, ranging from text classification to clustering and beyond. While these ML approaches typically yield favourable results from various languages, they may fall short when dealing with languages that have resource-intensive demands. In such cases, dedicated bilingual models, equipped with techniques like translation ranking tasks and trained on translation pairs, often outperform ML models in terms of producing closely aligned representations. Moreover, expanding ML models to encompass more languages, while maintaining high performance, can be a challenging endeavour due to limitations in model capacity and the occasionally suboptimal quality of training data for low-resource languages. Table 6.2 provides an overview of the hyperparameters specific to each language, offering insights into the fine-tuning process for optimal performance. The proposed approach incorporates a five-layered dense neural network, complete with batch normalization, as depicted in the block diagram presented in Figure 6.1.

Table 6.2 Hyperparameters of the 5-layered dense model

Languages	Arabic	Tamil	Vietnamese
Training samples	94	86	55
Testing samples	32	43	29
Dimensions	768 X 1	768 X 1	768 X 1
Epochs	100	100	100
Activation function	ReLu + Tanh	ReLu + Tanh	ReLu + Tanh
Number of Dense layers	5	5	5
With or without Batch Normalization	With	With	With

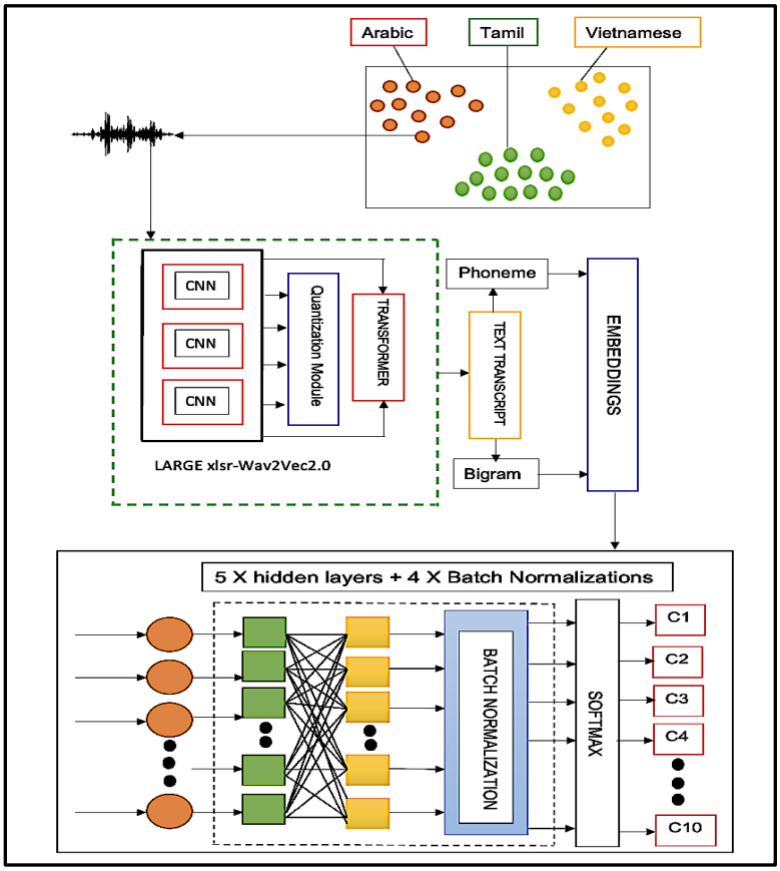


Figure 6.1. The block diagram of our proposed method

6.1.4 DEEP DENSE NEURAL NETWORK WITH BATCH NORMALIZATION

A DNN employs a basic nonlinear transformation to mathematically convert an input vector (X) into a set of feature maps. The input signals are denoted as an array $X = [X_1, X_2, X_3, ...,$

X_F]. These signals are processed by multiplying them with the corresponding synaptic weights, represented as elements in the array $W = [W_1, W_2, W_3, ..., W_F]$. This operation yields the value Z, often referred to as the "activation potential", as illustrated in the mathematical example provided in (6.1).

$$Z = \sum_{i=1}^{F} X_i W_i + b \tag{6.1}$$

In (6.1), the general operation of the dense layer is outlined with the assistance of the bias vectors denoted as 'b' and the weight matrices represented as 'W_i'.

$$Y_1^{(1)} = R(Z_1^{(1)}) , Y_2^{(1)} = R(Z_2^{(1)}) , \ldots\ldots , Y_F^{(1)} = R(Z_F^{(1)}) \tag{6.2}$$

$$Y_1^{(2)} = tanh(Z_1^{(2)}) , \; Y_2^{(2)} = \; tanh(Z_2^{(2)}) , \ldots\ldots , Y_E^{(2)} = \; tanh(Z_E^{(2)}) \tag{6.3}$$

The ReLU function is a non-linear activation function widely adopted in DL due to its specific advantages. Unlike some other activation methods, ReLU does not activate all neurons simultaneously. Its mathematical representation is expressed as '$R(X)$', and it can be defined as follows in (6.4):

$$R(X) = max\,(0, X) \tag{6.4}$$

In Equation 6.3, each element (1, 2, 3, and so on) corresponds to a different layer in the neural network, with each layer having a specific activation function. The neural network is structured as follows: the first layer employs the ReLU activation function, the second layer utilizes the *tanh* activation function, and the third layer again employs the ReLU activation function, and so forth. Batch normalization layers are inserted between each dense layer to enhance training stability. The subscripts "1" and "2" denote different sample instances within the network, while "E" signifies the final layer, which employs the sigmoid activation function. The notation "E" indicates the end of this final layer. The sigmoid function is represented by the symbol σ.

Historically, the tanh function gained more popularity than the sigmoid function, particularly for multi-layer neural networks. However, the vanishing gradient issue that affected sigmoids was not completely resolved by *tanh*. This problem has been more effectively addressed with the incorporation of ReLU activations.

$$tanh\,(x) = e^{x} - e^{-x} \;/\; e^{x} + e^{-x} \tag{6.5}$$

The sigmoid function is employed primarily because it outputs values within the range of 0 to 1. This characteristic makes it ideal for models focused on probability predictions. Given that probabilities naturally fall within the 0 to 1 interval, the sigmoid function is the most suitable choice for such applications.

$$\sigma = 1 \,/\, 1 + e^{-x} \tag{6.6}$$

As per the (6.4), $R(X)$ can be decomposed into its constituent components. To put it simply, this breakdown can be expressed as follows:

$$f(X) = \{0 \text{ for } X < 0, X \text{ for } X \geq 0\} \tag{6.7}$$

In order to mitigate the issue of internal covariate shift, the inputs of each layer undergo normalization using the batch normalization approach. During the training process, which typically involves SGD over randomized mini-batches of training data samples represented as $B \subset X$, batch normalization (BN) is commonly incorporated into modern DNN architectures. BN serves to centre and normalize the entries of feature maps using four additional parameters: μ_i, V_i, β_i, and γ_i (as introduced by Ioffe and Szegedy in 2015). The batch normalization layer fulfills the following functions during training:

It calculates the mean and variance of the input layers, as shown in (6.8). Once the input layer's mean and covariance have been determined, it normalizes the layer's inputs using the previously calculated batch statistics. After normalization, it employs scaling and shifting to produce the layer's output. Mathematically, the batch mean is represented as follows:

$$\mu = \frac{1}{F}\sum_{i=1}^{F} Xi \tag{6.8}$$

The mathematical representation for batch variance is denoted by σ^2

$$V^2 = \frac{1}{F}\sum_{i=1}^{F}(Xi - \mu) \tag{6.9}$$

Once you've calculated the mean and covariance of the input layers, you can proceed to normalize the layer inputs using the precomputed batch statistics.

$$\overline{Y}_i = X_{i-}\mu \,/\, \sqrt{sqrt(V) + \in} \tag{6.10}$$

In addition to the network's original parameters γ and β, during training, new parameters μ and σ are also learned. In the context of the BN algorithm, two key elements come into play: γ (gamma) and β (beta). These parameters serve distinct roles in the process, with γ being utilized for rescaling and β for shifting the vector containing values derived from preceding operations. These parameters μ and σ for each mini-batch B are computed directly using the mean and standard deviation of the current mini-batch feature maps. So, if there are F samples in each batch and B batches, the following formula is used to calculate the inference mean:

$$Y_i = \gamma \overline{X}_i + \beta \tag{6.11}$$

$$E_X = \frac{1}{F} \sum_{i=1}^{B} \mu^{(i)} \tag{6.12}$$

The formula provided below is used to calculate the inference variance

$$\text{Var}_x = \left(\frac{F}{F-1} \right) \frac{1}{F} \sum_{i=1}^{F} sqrt(V(i)) \tag{6.13}$$

Inference scaling is computed using the following formula:

$$Y = \gamma X / \sqrt{\text{Var } X + \in} + \left(\beta + \gamma \, E_X / \sqrt{\text{Var } X + \in} \right) \tag{6.14}$$

Batch normalization produces the output value Y, which is then fed into the neural network. During testing or inference, the mean and variance remain constant and are determined using the mean and variance values from previous training batches. Batch normalization acts as a straightforward linear transformation of the preceding layer's output, often a convolution, during inference. Its purpose is to enhance the partition areas around the training data, allowing for a more accurate approximation by adjusting the input space's spline partition to minimize the total least squares (TLS) distance between the spline partition boundaries and the layer's DN inputs.

$$\overline{Y} : \{ \overline{Y}_1 = \sigma(Z_1^{(N)}), \overline{Y}_2 = \sigma(Z_2^{(N)}), \overline{Y}_3 = \sigma(Z_3^{(N)}), \ldots\ldots, \overline{Y}_L = \sigma(Z_F^{(N)}) \} \tag{6.15}$$

P_{Class} should refer to the posterior class probability connected to the spoken word's class. We employ the functions outlined below to amalgamate the two probabilistic decision scores derived from the dense + batch normalization model. This model is trained on the fusion of the English LARGE xlsr-Wav2Vec2-53 text transcript converted to phonemes and the

English LARGE xlsr-Wav2Vec2-53 text transcript converted to bigrams. The final results of layer categorization are shown by the $\overline{Y}$. The test sample's class is determined using the formula below:

$$P_{Class} = \max (P_{Class}\overline{Y}) \quad (6.16)$$

$$\text{class} = \forall_c \, argmax \, (P_{Class}) \quad (6.17)$$

6.2 Experimental Results

6.2.1 Dataset

The multilingual spoken words Corpus (MSWC) (Mazumder et al. 2021) is a valuable resource used for both commercial applications like keyword and spoken phrase searches and academic research. It boasts a massive and continually growing audio dataset, encompassing more than 5 billion individuals who speak 50 different languages. This corpus, released under the CC-BY 4.0 license, is extensive, comprising 23.4 million 1-second spoken instances and over 340,000 keywords (equating to over 6,000 hours of audio). The applications of this dataset are diverse, ranging from automated contact centres to voice-activated consumer electronics. The dataset was created using forced alignment techniques on crowdsourced sentence-level audio, providing precise per-word time estimates for extraction. It is noteworthy that every alignment is meticulously included in the dataset.

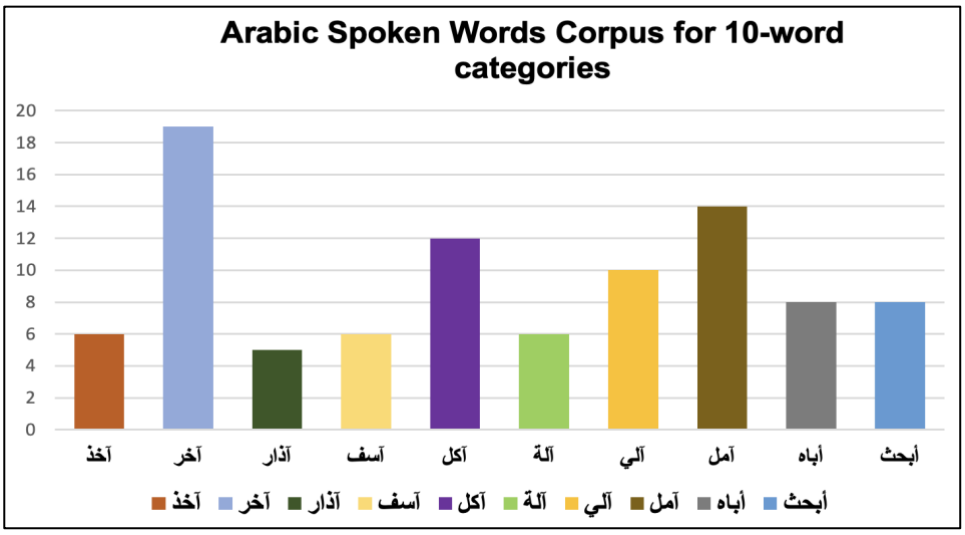


Figure 6.2. The number of samples in Arabic Multilingual Spoken Words Corpus

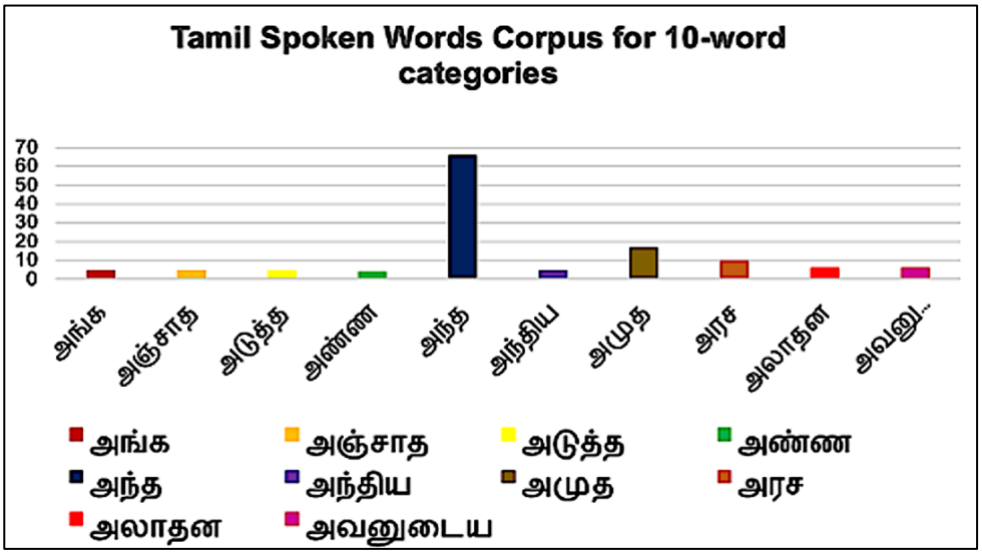


Figure 6.3. The number of samples in Tamil Multilingual Spoken Words Corpus

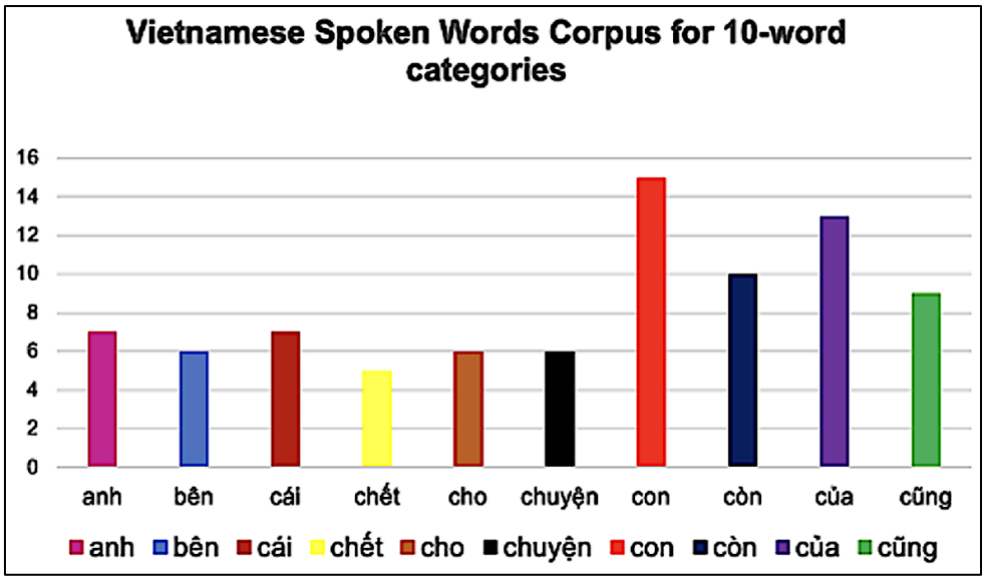


Figure 6.4. The number of samples in Vietnamese Multilingual Spoken Words Corpus

Furthermore, the dataset includes methods for identifying potential outliers and offers a comprehensive analysis of its contents. Additionally, it provides benchmark accuracy values for evaluating keyword detection algorithms in comparison to models trained on manually recorded keyword datasets. The total number of samples for each language is visually represented in Figure 6.2, 6.3, 6.4. The MSWC database comprises spoken words in a diverse set of 50 languages, encompassing both high-resource languages such as English and Spanish, and LRLs like Dhivehi and Oriya (an Indo-Aryan language spoken in the Indian

state of Odisha and in the Maldives). To classify the Arabic, Tamil, and Vietnamese languages, we customized the multilingual spoken words corpus, which was originally designed for various purposes such as academic research, commercial applications, keyword spotting, and spoken term search. However, we focused on a limited set of 10-word categories, resulting in a dataset with fewer audio samples. While this dataset is extensive, our specific adaptations for these 10 spoken word categories have created a significant bias. The MSWC collection encompasses spoken word audio from 26 low-resource languages, 12 medium-resource languages, and 12 high-resource languages, making it versatile for real-world applications, although it is predominantly utilized in academic research.

6.2.2 EXPERIMENTAL SETUP

The experiments were conducted using Python 3.10.4 on a Mac OS High Sierra system equipped with an Intel Core i5 CPU and Intel Graphics, operating at 1.8 GHz. We adopted a 70:30 train-to-test split ratio with 5-fold cross-validation for our research. Optimization was performed with the Adam optimizer, employing parameters such as a batch size of 16, 100 training epochs, and a learning rate of 0.01. In our experiments, we utilized both the "*ReLU*" and "*tanh*" activation functions. "*tanh*" showed superior performance for multi-layer neural networks, while "*ReLU*" introduced non-linearity. Therefore, we leveraged the strengths of both activation functions in our chapter.

6.2.3 RESULTS AND DISCUSSIONS

Developing speech recognition systems for languages with numerous phonemes and intricate stress and intonation patterns poses a significant challenge. Accents and dialects further complicate the accurate recognition of spoken words and phrases. However, Wav2Vec2.0 has proven to be a substantial improvement in speech recognition accuracy. This advancement can have considerable advantages for languages with complex phoneme systems and intonation patterns, such as Arabic, Tamil, and Vietnamese. Leveraging Wav2Vec2.0 for speech recognition in these languages can lead to more precise and efficient systems, better equipped to handle their distinctive characteristics. To process the English LARGE xlsr-Wav2Vec2-53 phonemes and bigrams, we transformed them into vectors. Before vectorization, we eliminated noisy phonemes and signal noise (silence) from the

collected phonemes. The phonemes were then converted into 768-dimensional vectors using language-agnostic embeddings. Although we initially attempted an early fusion of phone embeddings, the results were unsatisfactory, leading us to focus solely on late fusion. The five-layered dense model with batch normalization received the language-agnostic embeddings as input. The model's performance is at the forefront of various bi-text retrieval and mining tasks, offering extended language coverage. Each language's (Arabic, Tamil, and Vietnamese) feature matrix for each embedding is a 768 x 1 matrix. These matrices were input into the five-layered dense model with batch normalization, yielding superior results compared to current methods. The first layer of the dense model comprises 2048 neurons with ReLu activation. Subsequently, batch normalization is applied. The second layer consists of 1024 neurons with tanh activation, followed by batch normalization. The third layer includes 512 neurons with ReLu activation and, once again, batch normalization. The fourth layer features 256 neurons with tanh activation, followed by batch normalization. The fifth and final layer consists of 64 neurons with ReLu activation, and it culminates with ten neurons using the Softmax activation function, representing the ten spoken word categories. The confusion matrices for the fusion of phonemes and bigrams are illustrated in Figures 6.5 to 6.7.

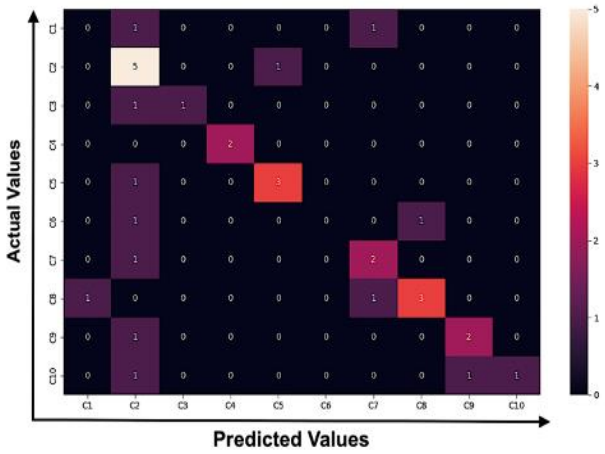


Figure 6.5. Confusion Matrix of Fusion Framework of Arabic Multilingual Spoken Words Corpus

Additionally, Tables 6.3 to 6.5 provide information on the classification of spoken words within specific languages. The evaluation of each language and category is presented through tables and figures, providing classification results. The assessment metrics for the various spoken word classes, including Sensitivity, Specificity, PPV, NPV, FPR, FNR, and

FDR, are outlined in Tables 6.6 to 6.8. While some classes exhibit high Sensitivity and PPV values, all classes show high Specificity and NPV values. The FPR values are consistently low for all classes, while FNR and FDR values vary.

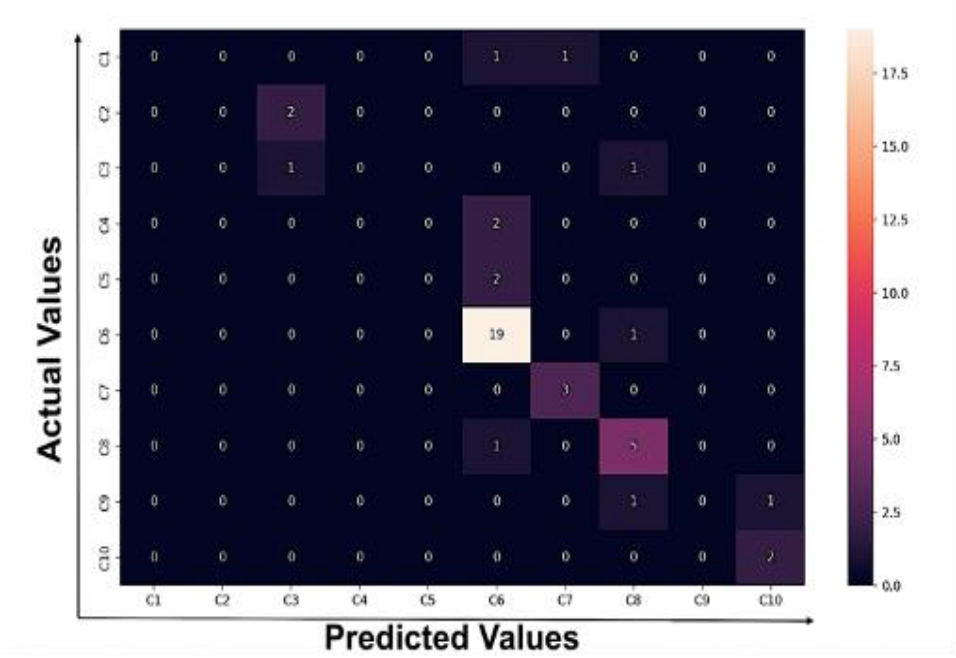


Figure 6.6. Confusion Matrix of Fusion Framework of Tamil Multilingual Spoken Words Corpus

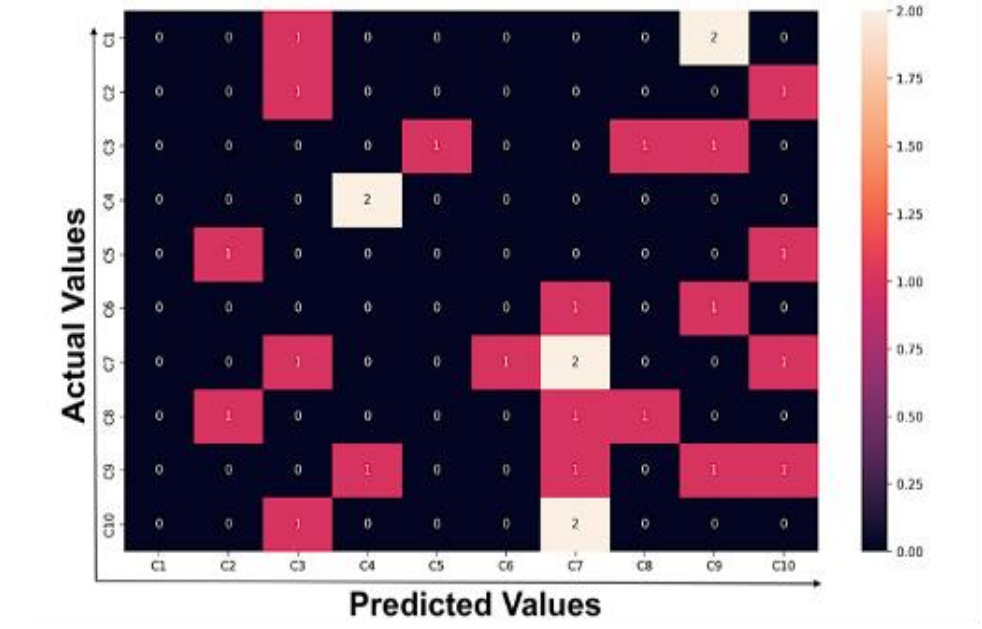


Figure 6.7. Confusion Matrix of Fusion Framework of Vietnamese Multilingual Spoken Words Corpus

Table 6.3 The classification results of fusion framework in Arabic Multilingual Spoken Words Corpus

Arabic Spoken word categories	Accuracy	WER
آخذ	1.00	0.00
آخر	1.00	0.00
آذار	0.93103448	0.07
آسف	0.89655172	0.10
آكل	0.89655172	0.10
آلة	0.96551724	0.03
آلي	0.89655172	0.10
آمل	0.86206897	0.14
أباه	0.89655172	0.10
أبحث	0.96551724	0.04

Table 6.4 The classification results of fusion framework in Vietnamese Multilingual Spoken Words Corpus

Vietnamese Spoken word categories	Accuracy	WER
anh	0.93103448	0.07
bên	0.86206897	0.14
cái	0.82758621	0.17
chết	0.93103448	0.07
cho	0.86206897	0.14
chuyện	0.89655172	0.11
con	0.79310345	0.21
còn	0.86206897	0.14
của	0.72413793	0.28
cũng	0.79310345	0.21

Table 6.5 The classification results of fusion framework in Tamil Multilingual Spoken Words Corpus

Tamil Spoken word categories	Accuracy	WER
அங்க	0.95348837	0.04
அஞ்சாத	0.95348837	0.04
அடுத்த	0.93023256	0.07
அண்ண	0.95348837	0.04
அந்த	0.95348837	0.04
அந்திய	0.8372093	0.16

அழுத	0.97674419	0.02
அரச	0.90697674	0.09
அலாதன	0.95348837	0.04
அவனுடைய	0.97674419	0.02

The provided Table (6.6 – 6.8) contains evaluation metrics for each word class, including accuracy, NPV, FPR, FNR, and FDR. NPV, or Negative Predictive Value, is a metric that assesses the proportion of cases with negative test results that are truly negative samples. It quantifies the percentage of subjects whose test results correctly identify them as negative cases out of all test-negative participants, including both true negatives and those incorrectly identified as true negatives. In an ideal test that doesn't produce any false negatives, the NPV value is 1, indicating 100% accuracy in identifying true negatives. Conversely, in a test that doesn't produce any true negatives, the NPV value is 0, meaning it fails to correctly identify any true negatives. For PPV, in a perfect test, the highest achievable result is 1 (100%), while the lowest possible value is 0. PPV represents the percentage of positive test results that correspond to true positives. The false negative rate (FNR), also referred to as the conditional likelihood of a negative test result given the presence of the positives being tested for, indicates the proportion of positive cases that are incorrectly identified as negative. The False Discovery Rate (FDR) quantifies the expected percentage of Type I errors. Table (6.9 – 6.11) presents evaluation metrics for spoken word categories in three languages, comparing our proposed technique (*) with seven other methods: (1) Zia and Zahid (2019), (2) Lezhenin et al. 2019, (3) Wazir et al. 2019, (4) Abdelmaksoud et al. 2021, (5) Haque et al. 2020, (6) Kherdekar and Naik, 2021, and (7) Passricha and Aggarwal, 2020.

Table 6.6 The classified results of fusion framework in Arabic MSWC compared with SOTA

Spoken word categories	Precision	Recall	f 1-Score	Sensitivity	Specificity	NPV	FPR	FNR	FDR	PPV
آخذ	0.20 (1) 0.00 (2) 0.00 (3) 0.25 (4) 0.00 (5) 0.25 (6) 0.25 (7) 0.00 (*)	0.50 (1) 0.00 (2) 0.00 (3) 0.12 (4) 0.00 (5) 0.10 (6) 0.10 (7) 0.00 (*)	0.29 (1) 0.00 (2) 0.00 (3) 0.17 (4) 0.00 (5) 0.14 (6) 0.14 (7) 0.00 (*)	0.50 (1) 0.00 (2) 0.00 (3) 0.13 (4) 0.00 (5) 0.10 (6) 0.10 (7) 0.00 (*)	0.85 (1) 0.93 (2) 0.92 (3) 0.95 (4) 1.00 (5) 0.95 (6) 0.95 (7) 1.00 (*)	0.96 (1) 0.96 (2) 0.88 (3) 0.89 (4) 0.86 (5) 0.85 (6) 0.85 (7) 0.94 (*)	0.15 (1) 0.08 (2) 0.08 (3) 0.05 (4) 0.00 (5) 0.05 (6) 0.05 (7) 0.00 (*)	0.50 (1) 1.00 (2) 1.00 (3) 0.88 (4) 1.00 (5) 0.90 (6) 0.90 (7) 1.00 (*)	0.80 (1) 1.00 (2) 1.00 (3) 0.75 (4) 1.00 (5) 0.75 (6) 0.75 (7) 1.00 (*)	0.20 (1) 0.00 (2) 0.00 (3) 0.25 (4) 0.00 (5) 0.25 (6) 0.25 (7) 0.00 (*)

آخر	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.79 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	1.00 (2)	0.90 (2)	0.00 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	0.97 (3)	0.94 (3)	0.03 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	0.97 (4)	0.97 (4)	0.03 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	1.00 (5)	0.94 (5)	0.00 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	1.00 (6)	0.50 (6)	0.67 (6)	0.50 (6)	1.00 (6)	0.97 (6)	0.00 (6)	0.50 (6)	0.00 (6)	1.00 (6)
	0.67 (7)	0.50 (7)	0.57 (7)	0.50 (7)	0.98 (7)	0.97 (7)	0.02 (7)	0.50 (7)	0.33 (7)	0.67 (7)
	0.45 (*)	0.83 (*)	0.59 (*)	0.83 (*)	0.77 (*)	0.95 (*)	0.23 (*)	0.17 (*)	0.55 (*)	0.45 (*)
آذار	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	0.89 (1)	0.92 (1)	0.11 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.33 (2)	0.50 (2)	0.40 (2)	0.50 (2)	0.73 (2)	0.85 (2)	0.26 (2)	0.50 (2)	0.67 (2)	0.00 (2)
	0.10 (3)	0.20 (3)	0.13 (3)	0.20 (3)	0.66 (3)	0.82 (3)	0.34 (3)	0.80 (3)	0.91 (3)	0.10 (3)
	0.21 (4)	0.50 (4)	0.29 (4)	0.50 (4)	0.57 (4)	0.84 (4)	0.43 (4)	0.50 (4)	0.79 (4)	0.21 (4)
	0.18 (5)	0.33 (5)	0.24 (5)	0.33 (5)	0.67 (5)	0.82 (5)	0.33 (5)	0.67 (5)	0.82 (5)	0.18 (5)
	0.17 (6)	0.29 (6)	0.21 (6)	0.28 (6)	0.83 (6)	0.91 (6)	0.17 (6)	0.71 (6)	0.83 (6)	0.17 (6)
	0.17 (7)	0.29 (7)	0.21 (7)	0.28 (7)	0.83 (7)	0.91 (7)	0.17 (7)	0.71 (7)	0.83 (7)	0.17 (7)
	0.50 (*)	0.50 (*)	0.50 (*)	0.50 (*)	0.97 (*)	0.97 (*)	0.03 (*)	0.50 (*)	0.50 (*)	0.50 (*)
آسف	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.86 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	1.00 (2)	0.93 (2)	0.00 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	0.98 (3)	0.89 (3)	0.02 (3)	1.00 (3)	1.00 (3)	0.10 (3)
	0.14 (4)	0.25 (4)	0.18 (4)	0.25 (4)	0.90 (4)	0.95 (4)	0.10 (4)	0.75 (4)	0.86 (4)	0.14 (4)
	0.14 (5)	0.80 (5)	0.24 (5)	0.80 (5)	0.59 (5)	0.97 (5)	0.41 (5)	0.20 (5)	0.87 (5)	0.14 (5)
	0.60 (6)	0.50 (6)	0.55 (6)	0.50 (6)	0.97 (6)	0.95 (6)	0.03 (6)	0.50 (6)	0.40 (6)	0.60 (6)
	0.60 (7)	0.50 (7)	0.55(7)	0.50 (7)	0.97 (7)	0.95 (7)	0.03 (7)	0.50 (7)	0.40 (7)	0.60 (7)
	0.67 (*)	1.00 (*)	0.80 (*)	1.00 (*)	0.97 (*)	1.00 (*)	0.03 (*)	0.00 (*)	0.33 (*)	0.67 (*)
أكل	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.90 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	1.00 (2)	0.96 (2)	0.00 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	0.94 (3)	0.95 (3)	0.06 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	0.98 (4)	0.94 (4)	0.02 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	0.90 (5)	0.91 (5)	0.10 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	0.00 (6)	0.00 (6)	0.00 (6)	0.00 (6)	1.00 (6)	0.92 (6)	0.00 (6)	1.00 (6)	1.00 (6)	0.00 (6)
	1.00 (7)	0.20 (7)	0.33 (7)	0.20 (7)	1.00 (7)	0.94 (7)	0.00 (7)	0.80 (7)	0.00 (7)	1.00 (7)
	0.80 (*)	1.00 (*)	0.89 (*)	1.00 (*)	0.97 (*)	1.00 (*)	0.04 (*)	0.00 (*)	0.20 (*)	0.80 (*)
آلة	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.96 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	1.00 (2)	0.93 (2)	0.00 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	0.93 (3)	0.92 (3)	0.06 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	1.00 (4)	0.94 (4)	0.00 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	1.00 (5)	0.94 (5)	0.00 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	1.00 (6)	0.33 (6)	0.50 (6)	0.33 (6)	1.00 (6)	0.97 (6)	0.00 (6)	0.67 (6)	1.00 (6)	1.00 (6)
	1.00 (7)	0.33 (7)	0.50 (7)	0.33 (7)	1.00 (7)	0.97 (7)	0.00 (7)	0.67 (7)	0.00 (7)	1.00 (7)
	0.50 (*)	0.50 (*)	0.50 (*)	0.50 (*)	0.97 (*)	0.96 (*)	0.03 (*)	0.50 (*)	0.50 (*)	0.50 (*)
آلي	0.50 (1)	0.50 (1)	0.50 (1)	0.50 (1)	0.96 (1)	0.96 (1)	0.04 (1)	0.50 (1)	0.50 (1)	0.50 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	0.89 (2)	0.92 (2)	0.92 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	1.00 (3)	0.94 (3)	0.00 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	1.00 (4)	0.93 (4)	0.00 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	0.97 (5)	0.92 (5)	0.03 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	0.22 (6)	0.33 (6)	0.27 (6)	0.33 (6)	0.88 (6)	0.93 (6)	0.11 (6)	0.67 (6)	0.78 (6)	0.22 (6)
	0.22 (7)	0.33 (7)	0.27 (7)	0.33 (7)	0.88 (7)	0.93 (7)	0.12 (7)	0.67 (7)	0.78 (7)	0.22 (7)
	0.50 (*)	0.67 (*)	0.57 (*)	0.67 (*)	0.93 (*)	0.96 (*)	0.07 (*)	0.33 (*)	0.50 (*)	0.50 (*)
آمل	0.33 (1)	0.67 (1)	0.44 (1)	0.67 (1)	0.65 (1)	0.88 (1)	0.35 (1)	0.33 (1)	0.67 (1)	0.33 (1)
	0.33 (2)	0.50 (2)	0.40 (2)	0.50 (2)	0.74 (2)	0.85 (2)	0.85 (2)	0.50 (2)	0.67 (2)	0.33 (2)
	0.21 (3)	0.46 (3)	0.29 (3)	0.00 (3)	0.56 (3)	0.81 (3)	0.43 (3)	0.54 (3)	0.54 (3)	0.21 (3)
	0.29 (4)	0.40 (4)	0.33 (4)	0.04 (4)	0.71 (4)	0.80 (4)	0.29 (4)	0.60 (4)	0.71 (4)	0.29 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	1.00 (5)	0.80 (5)	0.00 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	0.19 (6)	0.38 (6)	0.25 (6)	0.38 (6)	0.58 (6)	0.79 (6)	0.41 (6)	0.62 (6)	0.81 (6)	0.19 (6)
	0.20 (7)	0.38 (7)	0.26 (7)	0.38 (7)	0.62 (7)	0.80 (7)	0.38 (7)	0.62 (7)	0.80 (7)	0.20 (7)
	1.00 (*)	0.60 (*)	0.75 (*)	0.60 (*)	1.00 (*)	0.93 (*)	0.00 (*)	0.40 (*)	0.00 (*)	1.00 (*)
أباه	0.17 (1)	0.50 (1)	0.25 (1)	0.50 (1)	0.81 (1)	0.96 (1)	0.19 (1)	0.50 (1)	0.83 (1)	0.17 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	0.77 (2)	0.87 (2)	0.87 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.46 (3)	1.00 (3)	0.88 (3)	0.00 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	1.00 (4)	0.89 (4)	0.00 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	0.93 (5)	0.92 (5)	0.06 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	0.33 (6)	0.25 (6)	0.29 (6)	0.25 (6)	0.93 (6)	0.90 (6)	0.07 (6)	0.75 (6)	0.67 (6)	0.33 (6)
	0.33 (7)	0.25 (7)	0.29 (7)	0.25 (7)	0.93 (7)	0.90 (7)	0.07 (7)	0.75 (7)	0.67 (7)	0.33 (7)
	1.00 (*)	0.33 (*)	0.50 (*)	0.33 (*)	1.00 (*)	0.94 (*)	0.00 (*)	0.67 (*)	0.00 (*)	1.00 (*)
أبحث	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	0.96 (1)	0.96 (1)	0.03 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	1.00 (2)	0.89 (2)	0.90 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	1.00 (3)	0.92 (3)	0.00 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	0.97 (4)	0.92 (4)	0.03 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	0.95 (5)	0.94 (5)	0.05 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	0.00 (6)	0.00 (6)	0.00 (6)	0.00 (6)	1.00 (6)	0.94 (6)	0.00 (6)	1.00 (6)	1.00 (6)	0.00 (6)
	0.00 (7)	0.00 (7)	0.00 (7)	0.00 (7)	1.00 (7)	0.94 (7)	0.00 (7)	1.00 (7)	1.00 (7)	0.00 (7)
	0.00 (*)	0.00 (*)	0.00 (*)	0.00 (*)	0.97 (*)	0.90 (*)	0.03 (*)	1.00 (*)	1.00 (*)	0.00 (*)

Table 6.7 The classified results of fusion framework in Tamil MSWC compared with SOTA

Spoken word categories	Precision	Recall	f 1-Score	Sensitivity	Specificity	NPV	FPR	FNR	FDR	PPV
அங்க	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.95 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	1.00 (2)	0.95 (2)	0.00 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	1.00 (3)	0.95 (3)	0.00 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	1.00 (4)	0.95 (4)	0.00 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	1.00 (5)	0.95 (5)	0.00 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	0.00 (6)	0.00 (6)	0.00 (6)	0.00 (6)	1.00 (6)	0.95 (6)	0.00 (6)	1.00 (6)	1.00 (6)	0.00 (6)
	0.00 (7)	0.00 (7)	0.00 (7)	0.00 (7)	1.00 (7)	0.95 (7)	0.00 (7)	1.00 (7)	1.00 (7)	0.00 (7)
	0.00 (*)	0.00 (*)	0.00 (*)	0.00 (*)	1.00 (*)	0.95 (*)	0.00 (*)	1.00 (*)	1.00 (*)	0.00 (*)
அஞ்சாத	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.95 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	1.00 (2)	0.95 (2)	0.00 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	1.00 (3)	0.50 (3)	0.67 (3)	0.50 (3)	1.00 (3)	0.97 (3)	0.00 (3)	0.50 (3)	0.00 (3)	1.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	1.00 (4)	0.95 (4)	0.00 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	1.00 (5)	0.50 (5)	0.67 (5)	0.50 (5)	1.00 (5)	0.98 (5)	0.00 (5)	0.50 (5)	0.00 (5)	1.00 (5)
	0.00 (6)	0.00 (6)	0.00 (6)	0.00 (6)	0.98 (6)	0.95 (6)	0.02 (6)	1.00 (6)	1.00 (6)	0.00 (6)
	0.33 (7)	0.50 (7)	0.40 (7)	0.50 (7)	0.95 (7)	0.97 (7)	0.05 (7)	0.50 (7)	0.67 (7)	0.33 (7)
	0.00 (*)	0.00 (*)	0.00 (*)	0.00 (*)	1.00 (*)	0.95 (*)	0.00 (*)	1.00 (*)	1.00 (*)	0.00 (*)
அடுத்த	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.95 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	1.00 (2)	0.95 (2)	0.00 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	1.00 (3)	0.95 (3)	0.00 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	1.00 (4)	0.50 (4)	0.67 (4)	0.50 (4)	1.00 (4)	0.98 (4)	0.00 (4)	0.50 (4)	0.00 (4)	1.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	1.00 (5)	0.95 (5)	0.00 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	1.00 (6)	0.50 (6)	0.67 (6)	0.50 (6)	1.00 (6)	0.97 (6)	0.00 (6)	0.50 (6)	0.00 (6)	1.00 (6)
	0.00 (7)	0.00 (7)	0.00 (7)	0.00 (7)	1.00 (7)	0.95 (7)	0.00 (7)	1.00 (7)	1.00 (7)	0.00 (7)
	0.33 (*)	0.50 (*)	0.40 (*)	0.50 (*)	0.95 (*)	0.97 (*)	0.05 (*)	0.50 (*)	0.67 (*)	0.33 (*)
அண்ண	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.95 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	1.00 (2)	0.95 (2)	0.00 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	1.00 (3)	0.95 (3)	0.00 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	1.00 (4)	0.95 (4)	0.00 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	1.00 (5)	0.95 (5)	0.00 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	0.00 (6)	0.00 (6)	0.00 (6)	0.00 (6)	1.00 (6)	0.95 (6)	0.00 (6)	1.00 (6)	1.00 (6)	0.00 (6)
	0.00 (7)	0.00 (7)	0.00 (7)	0.00 (7)	1.00 (7)	0.95 (7)	0.00 (7)	1.00 (7)	1.00 (7)	0.00 (7)
	0.00 (*)	0.00 (*)	0.00 (*)	0.00 (*)	1.00 (*)	0.95 (*)	0.00 (*)	1.00 (*)	1.00 (*)	0.00 (*)
அந்த	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.95 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.11 (2)	0.50 (2)	0.18 (2)	0.50 (2)	0.80 (2)	0.97 (2)	0.20 (2)	0.50 (2)	0.88 (2)	0.11 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	0.85 (3)	0.94 (3)	0.15 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	0.98 (4)	0.95 (4)	0.03 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	0.85 (5)	0.94 (5)	0.15 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	0.00 (6)	0.00 (6)	0.00 (6)	0.00 (6)	0.90 (6)	0.95 (6)	0.10 (6)	1.00 (6)	1.00 (6)	0.00 (6)
	0.00 (7)	0.00 (7)	0.00 (7)	0.00 (7)	0.95 (7)	0.95 (7)	0.05 (7)	1.00 (7)	1.00 (7)	0.00 (7)
	0.00 (*)	0.00 (*)	0.00 (*)	0.00 (*)	1.00 (*)	0.95 (*)	0.00 (*)	1.00 (*)	1.00 (*)	0.00 (*)
அந்திய	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.95 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	0.98 (2)	0.95 (2)	0.03 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	1.00 (3)	0.95 (3)	0.00 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	1.00 (4)	0.95 (4)	0.00 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	1.00 (5)	0.95 (5)	0.00 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	0.00 (6)	0.00 (6)	0.00 (6)	0.00 (6)	1.00 (6)	0.95 (6)	0.00 (6)	1.00 (6)	1.00 (6)	0.00 (6)
	0.00 (7)	0.00 (7)	0.00 (7)	0.00 (7)	1.00 (7)	0.95 (7)	0.00 (7)	1.00 (7)	1.00 (7)	0.00 (7)
	0.76 (*)	0.95 (*)	0.84 (*)	0.95 (*)	0.74 (*)	0.94 (*)	0.26 (*)	0.05 (*)	0.05 (*)	0.76 (*)
அமுத	0.45 (1)	1.00 (1)	0.62 (1)	1.00 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)	0.55 (1)	0.45 (1)
	0.53 (2)	0.89 (2)	0.67 (2)	0.89 (2)	0.35 (2)	0.80 (2)	0.65 (2)	0.11 (2)	0.47 (2)	0.53 (2)
	0.58 (3)	1.00 (3)	0.73 (3)	1.00 (3)	0.39 (3)	1.00 (3)	0.61 (3)	0.00 (3)	0.42 (3)	0.58 (3)
	0.50 (4)	1.00 (4)	0.67 (4)	1.00 (4)	0.17 (4)	1.00 (4)	0.83 (4)	0.00 (4)	0.50 (4)	0.50 (4)
	0.55 (5)	0.95 (5)	0.69 (5)	0.95 (4)	0.35 (5)	0.89 (5)	0.65 (5)	0.05 (5)	0.45 (5)	0.55 (5)
	0.56 (6)	1.00 (6)	0.72 (6)	0.72 (6)	0.34 (6)	1.00 (6)	0.65 (6)	0.00 (6)	0.44 (6)	0.56 (6)
	0.56 (7)	0.95 (7)	0.71 (7)	0.95 (7)	0.39 (7)	0.90 (7)	0.60 (7)	0.05 (7)	0.05 (7)	0.56 (7)
	0.75 (*)	1.00 (*)	0.86 (*)	1.00 (*)	0.98 (*)	1.00 (*)	0.03 (*)	0.00 (*)	0.00 (*)	0.75 (*)
அரச	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.95 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	1.00 (2)	0.93 (2)	0.00 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	1.00 (3)	0.93 (3)	0.00 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	1.00 (4)	0.93 (4)	0.00 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	0.97 (5)	0.93 (5)	0.02 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	0.00 (6)	0.00 (6)	0.00 (6)	0.00 (6)	1.00 (6)	0.93 (6)	0.00 (6)	1.00 (6)	1.00 (6)	0.00 (6)
	0.00 (7)	0.00 (7)	0.00 (7)	0.00 (7)	0.95 (7)	0.93 (7)	0.05 (7)	1.00 (7)	1.00 (7)	0.00 (7)
	0.62 (*)	0.83 (*)	0.71 (*)	0.83 (*)	0.92 (*)	0.97 (*)	0.08 (*)	0.17 (*)	0.17 (*)	0.62 (*)
அலாதன	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	1.00 (1)	0.95 (1)	0.00 (1)	1.00 (1)	1.00 (1)	0.00 (1)
	0.00 (2)	0.00 (2)	0.00 (2)	0.00 (2)	1.00 (2)	0.95 (2)	0.00 (2)	1.00 (2)	1.00 (2)	0.00 (2)
	0.00 (3)	0.00 (3)	0.00 (3)	0.00 (3)	1.00 (3)	0.95 (3)	0.00 (3)	1.00 (3)	1.00 (3)	0.00 (3)
	0.00 (4)	0.00 (4)	0.00 (4)	0.00 (4)	1.00 (4)	0.95 (4)	0.00 (4)	1.00 (4)	1.00 (4)	0.00 (4)
	0.00 (5)	0.00 (5)	0.00 (5)	0.00 (5)	1.00 (5)	0.95 (5)	0.00 (5)	1.00 (5)	1.00 (5)	0.00 (5)
	0.00 (6)	0.00 (6)	0.00 (6)	0.00 (6)	0.98 (6)	0.95 (6)	0.03 (6)	1.00 (6)	1.00 (6)	0.00 (6)
	0.00 (7)	0.00 (7)	0.00 (7)	0.00 (7)	1.00 (7)	0.95 (7)	0.00 (7)	1.00 (7)	1.00 (7)	0.00 (7)
	0.00 (*)	0.00 (*)	0.00 (*)	0.00 (*)	1.00 (*)	0.95 (*)	0.00 (*)	1.00 (*)	1.00 (*)	0.00 (*)

அவனுடைய	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.67 (7) 0.67 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.33 (7) 1.00 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.44 (7) 0.80 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.33 (7) 1.00 (*)	1.00 (1) 1.00 (2) 0.94 (3) 0.94 (4) 0.97 (5) 0.97 (6) 0.97 (6) 0.98 (*)	0.86 (1) 0.86 (2) 0.85 (3) 0.85 (4) 0.85 (5) 0.85 (6) 0.90 (7) 1.00 (*)	0.00 (1) 0.00 (2) 0.06 (3) 0.06 (4) 0.03 (5) 0.03 (6) 0.03 (7) 0.02 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 0.67 (7) 0.00 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 0.33 (7) 0.33 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.67 (7) 0.67 (*)

Table 6.8 The classified results of fusion framework in Vietnamese MSWC compared with SOTA

Spoken word categories	Precision	Recall	f 1-Score	Sensitivity	Specificity	NPV	FPR	FNR	FDR	PPV
anh	0.00 (1) 1.00 (2) 0.00 (3) 0.00 (4) 1.00 (5) 0.08 (6) 0.00 (7) 0.00 (*)	0.00 (1) 0.50 (2) 0.00 (3) 0.00 (4) 0.50 (5) 1.00 (6) 0.00 (7) 0.00 (*)	0.00 (1) 0.67 (2) 0.00 (3) 0.00 (4) 0.67 (5) 0.15 (6) 0.00 (7) 0.00 (*)	0.00 (1) 0.50 (2) 0.00 (3) 0.00 (4) 0.50 (5) 1.00 (6) 0.00 (7) 0.00 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 0.19 (6) 1.00 (7) 1.00 (*)	0.93 (1) 0.96 (2) 0.93 (3) 0.93 (4) 0.96 (5) 1.00 (6) 0.93 (7) 0.93 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.81 (6) 0.00 (7) 0.00 (*)	1.00 (1) 0.50 (2) 1.00 (3) 1.00 (4) 0.50 (5) 0.00 (6) 1.00 (7) 1.00 (*)	1.00 (1) 0.00 (2) 1.00 (3) 1.00 (4) 0.00 (5) 0.92 (6) 1.00 (7) 1.00 (*)	0.00 (1) 1.00 (2) 0.00 (3) 0.00 (4) 1.00 (5) 0.08 (6) 0.00 (7) 0.00 (*)
bên	0.06 (1) 0.00 (2) 0.14 (3) 0.50 (4) 0.05 (5) 0.00 (6) 0.00 (7) 0.00 (*)	0.50 (1) 0.00 (2) 0.50 (3) 0.50 (4) 0.50 (5) 0.00 (6) 0.00 (7) 0.00 (*)	0.11 (1) 0.00 (2) 0.22 (3) 0.09 (4) 0.09 (5) 0.00 (6) 0.00 (7) 0.00 (*)	0.50 (1) 0.00 (2) 0.50 (3) 0.50 (4) 0.50 (5) 0.00 (6) 0.00 (7) 0.00 (*)	0.41 (1) 1.00 (2) 0.78 (3) 0.30 (4) 0.29 (5) 1.00 (6) 0.78 (7) 0.93 (*)	0.92 (1) 0.93 (2) 0.95 (3) 0.89 (4) 0.89 (5) 0.93 (6) 0.92 (7) 0.93 (*)	0.59 (1) 0.00 (2) 0.22 (3) 0.70 (4) 0.70 (5) 0.00 (6) 0.22 (7) 0.07 (*)	0.50 (1) 1.00 (2) 0.50 (3) 0.50 (4) 0.50 (5) 1.00 (6) 1.00 (7) 1.00 (*)	0.94 (1) 1.00 (2) 0.86 (3) 0.95 (4) 0.95 (5) 1.00 (6) 1.00 (7) 1.00 (*)	0.06 (1) 0.00 (2) 0.14 (3) 0.50 (4) 0.05 (5) 0.00 (6) 0.00 (7) 0.00 (*)
cái	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.25 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.33 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.29 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.33 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 1.00 (7) 0.88 (*)	0.90 (1) 0.90 (2) 0.90 (3) 0.90 (4) 0.90 (5) 0.90 (6) 0.90 (7) 0.92 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.12 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 1.00 (7) 0.67 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 1.00 (7) 0.75 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.25 (*)
chết	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.67 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.67 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.67 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.67 (*)	1.00 (1) 1.00 (2) 0.92 (3) 1.00 (4) 1.00 (5) 1.00 (6) 0.96 (7) 0.96 (*)	0.90 (1) 0.90 (2) 0.89 (3) 0.90 (4) 0.90 (5) 0.90 (6) 0.89 (7) 0.96 (*)	0.00 (1) 0.00 (2) 0.08 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.03 (7) 0.04 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 1.00 (7) 0.33 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 1.00 (7) 0.33 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.67 (*)
cho	0.08 (1) 0.00 (2) 0.00 (3) 0.25 (4) 0.33 (5) 0.00 (6) 0.11 (7) 0.00 (*)	0.33 (1) 0.00 (2) 0.00 (3) 0.33 (4) 0.33 (5) 0.00 (6) 0.33 (7) 0.00 (*)	0.13 (1) 0.00 (2) 0.00 (3) 0.29 (4) 0.33 (5) 0.00 (6) 0.17 (7) 0.00 (*)	0.33 (1) 0.00 (2) 0.00 (3) 0.33 (4) 0.33 (5) 0.00 (6) 0.33 (7) 0.00 (*)	0.57 (1) 0.96 (2) 0.92 (3) 0.88 (4) 0.92 (5) 0.80 (6) 0.69 (7) 0.96 (*)	0.88 (1) 0.90 (2) 0.89 (3) 0.92 (4) 0.92 (5) 0.88 (6) 0.90 (7) 0.90 (*)	0.42 (1) 0.04 (2) 0.08 (3) 0.12 (4) 0.07 (5) 0.19 (6) 0.30 (7) 0.04 (*)	0.67 (1) 1.00 (2) 1.00 (3) 0.67 (4) 0.67 (5) 1.00 (6) 0.67 (7) 1.00 (*)	0.92 (1) 1.00 (2) 1.00 (3) 0.75 (4) 0.67 (5) 1.00 (6) 0.89 (7) 1.00 (*)	0.08 (1) 0.00 (2) 0.00 (3) 0.25 (4) 0.33 (5) 0.00 (6) 0.11 (7) 0.00 (*)
chuyện	0.00 (1) 1.00 (2) 0.17 (3) 1.00 (4) 1.00 (5) 0.00 (6) 0.00 (7) 0.00 (*)	0.00 (1) 0.50 (2) 0.50 (3) 0.50 (4) 0.50 (5) 0.00 (6) 0.00 (7) 0.00 (*)	0.00 (1) 0.67 (2) 0.25 (3) 0.67 (4) 0.67 (5) 0.00 (6) 0.00 (7) 0.00 (*)	0.00 (1) 0.50 (2) 0.50 (3) 0.50 (4) 0.50 (5) 0.00 (6) 0.00 (7) 0.00 (*)	1.00 (1) 1.00 (2) 0.81 (3) 1.00 (4) 1.00 (5) 1.00 (6) 0.96 (7) 0.96 (*)	0.93 (1) 0.96 (2) 0.95 (3) 0.96 (4) 0.96 (5) 0.93 (6) 0.93 (7) 0.93 (*)	0.00 (1) 0.00 (2) 0.18 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.04 (7) 0.04 (*)	1.00 (1) 0.50 (2) 0.50 (3) 0.50 (4) 0.50 (5) 1.00 (6) 1.00 (7) 1.00 (*)	1.00 (1) 0.00 (2) 0.83 (3) 0.00 (4) 0.00 (5) 1.00 (6) 1.00 (7) 1.00 (*)	0.00 (1) 1.00 (2) 0.17 (3) 1.00 (4) 1.00 (5) 0.00 (6) 0.00 (7) 0.00 (*)
con	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.29 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.67 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.40 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.67 (*)	1.00 (1) 0.92 (2) 1.00 (3) 0.85 (4) 0.85 (5) 1.00 (6) 0.92 (7) 0.81 (*)	0.90 (1) 0.88 (2) 0.90 (3) 0.88 (4) 0.88 (5) 0.90 (6) 0.88 (7) 0.95 (*)	0.00 (1) 0.08 (2) 0.00 (3) 0.15 (4) 0.15 (5) 0.00 (6) 0.08 (7) 0.19 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 1.00 (7) 0.33 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 1.00 (7) 0.71 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.29 (*)
còn	0.00 (1) 0.10 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6)	0.00 (1) 1.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6)	0.00 (1) 0.17 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6)	0.00 (1) 1.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6)	1.00 (1) 0.30 (2) 0.59 (3) 1.00 (4) 1.00 (5) 1.00 (6)	0.93 (1) 1.00 (2) 0.89 (3) 0.93 (4) 0.93 (5) 0.93 (6)	0.00 (1) 0.70 (2) 0.40 (3) 0.00 (4) 0.00 (5) 0.00 (6)	1.00 (1) 0.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6)	1.00 (1) 0.90 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6)	0.00 (1) 0.10 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6)

	0.00 (7) 0.00 (*)	0.00 (7) 0.00 (*)	0.00 (7) 0.00 (*)	0.00 (7) 0.00 (*)	1.00 (7) 0.93 (*)	0.93 (7) 0.93 (*)	0.00 (7) 0.07 (*)	1.00 (7) 1.00 (*)	1.00 (7) 1.00 (*)	0.00 (7) 0.00 (*)
của	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.20 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.20 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.20 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.20 (*)	1.00 (1) 1.00 (2) 0.96 (3) 1.00 (4) 1.00 (5) 1.00 (6) 1.00 (7) 0.83 (*)	0.83 (1) 0.83 (2) 0.82 (3) 0.83 (4) 0.83 (5) 0.83 (6) 0.83 (7) 0.83 (*)	0.00 (1) 0.00 (2) 0.04 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.17 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 1.00 (7) 0.80 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 1.00 (7) 0.80 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.00 (7) 0.20 (*)
cũng	0.00 (1) 0.33 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.10 (7) 0.25 (*)	0.00 (1) 0.25 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.25 (7) 0.25 (*)	0.00 (1) 0.29 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.14 (7) 0.25 (*)	0.00 (1) 0.25 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.25 (7) 0.25 (*)	1.00 (1) 0.92 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 0.64 (7) 0.88 (*)	0.86 (1) 0.88 (2) 0.86 (3) 0.86 (4) 0.86 (5) 0.86 (6) 0.84 (7) 0.88 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.36 (7) 0.12 (*)	1.00 (1) 0.75 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 0.75 (7) 0.75 (*)	1.00 (1) 0.67 (2) 1.00 (3) 1.00 (4) 1.00 (5) 1.00 (6) 0.90 (7) 0.75 (*)	0.00 (1) 0.33 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.00 (6) 0.10 (7) 0.25 (*)

$$WER = 1 - ACC \tag{6.18}$$

We compared our proposed approach with several alternative methodologies, and the effectiveness of our late fusion model is showcased in Table (6.9 – 6.11). A CNN utilizing MFCC as input features (Haque et al. 2020) was outperformed by our technique in Arabic by 47.26%, Vietnamese by 10.34%, and Tamil by 24.54%. In sound categorization, it was observed that CNN with Mel Spectrogram as input outperforms LSTM (Lezhenin et al. 2019). Our method outperforms Mel Spectrogram with LSTM, achieving results of 38.69% in Arabic, 26.91% in Tamil, and 6.89% in Vietnamese. RNNs have made significant progress in audio modeling, but they haven't been widely utilized for Urdu acoustics due to vocabulary limitations and high computational costs. Even modern LSTM and GRU models (Zia and Zahid et al. 2019) did not surpass our approach. GFCC (Abdelmaksoud et al. 2021) and MFCC with CNN (Haque et al. 2020; Kherdekar and Naik, 2021) were found to be less effective with the MSWC than our method, surpassing even hybrid CNN and Bi-directional Long Short-Term Memory (BiLSTM) techniques. Table (6.9 – 6.11) provides evaluation metrics for spoken word categories across three languages and seven comparative methods: (1) Zia and Zahid et al. 2019, (2) Lezhenin et al. 2019, (3) Wazir et al. 2019, (4) Abdelmaksoud et al. 2021, (5) Haque et al. 2020, (6) Kherdekar and Naik, 2021, (7) Passricha and Aggarwal, 2020, and (*) our proposed technique. Our late fusion of phone embeddings and bigrams achieves an accuracy of 59.38% (Arabic), 24.13% (Vietnamese), and 69.77% (Tamil) for the 10 spoken word categories in the multilingual spoken words corpus, surpassing existing methods. This research highlights the potential of linguistics (phonemes and morphemes) in SWR, even with limited and imbalanced training samples.

By utilizing text-based embeddings and ML sentence embeddings techniques, we can achieve high accuracy levels comparable to cutting-edge methods.

Table 6.9 The classification results of fusion framework in Arabic MSWC compared with SOTA

Comparison Method	Accuracy	MCC	Macro	Micro
MFCC + BiLSTM + GRU (Zia and Zahid et al. 2019)	24.14%	0.1458	0.1480	0.2413
Mel-Spectrogram + BiLSTM (Lezhenin et al. 2019)	20.69%	0.0590	0.0800	0.2068
MFCC + LSTM (Wazir et al. 2019)	12.12 %	-0.0414	0.0414	0.1212
GFCC + CNN (Abdelmaksoud et al. 2021)	21.21%	0.0544	0.0974	0.2121
MFCC + CNN (Haque et al. 2020)	12.12%	0.0139	0.0470	0.1212
MFCC + CNN (Kherdekar and Naik, 2021)	27.27%	0.1597	0.2867	0.2727
Hybrid CNN + BiLSTM (Passricha and Aggarwal, 2020)	28.78%	0.1803	0.3119	0.2878
Proposed Method (*)	59.38%	0.5470	0.5098	0.5937

Table 6.10 The classification results of fusion framework in Tamil MSWC Corpus compared with SOTA

Comparison Method	Accuracy	MCC	Macro	Micro
MFCC + BiLSTM + GRU (Zia and Zahid et al. 2019)	45.23%	0.0000	0.0622	0.4523
Mel-Spectrogram + BiLSTM (Lezhenin et al. 2019)	42.857%	0.1368	0.0848	0.4285
MFCC + LSTM (Wazir et al. 2019)	47.61%	0.2036	0.1397	0.4761
GFCC + CNN (Abdelmaksoud et al. 2021)	47.61%	0.1578	0.1333	0.4761
MFCC + CNN (Haque et al. 2020)	45.23%	0.1609	0.1358	0.4523
MFCC + CNN (Kherdekar and Naik, 2021)	47.61%	0.1967	0.1383	0.4761
Hybrid CNN + BiLSTM (Passricha and Aggarwal, 2020)	50.00%	0.2462	0.1550	0.5000
Proposed Method (*)	69.77%	0.5768	0.3615	0.6976

Table 6.11 The classification results of fusion framework in Vietnamese MSWC compared with SOTA

COMPARISON METHOD	Accuracy	MCC	Macro	Micro
MFCC + BiLSTM + GRU (Zia and Zahid et al. 2019)	6.89%	-0.0217	0.0238	0.0689
Mel-Spectrogram + BiLSTM (Lezhenin et al. 2019)	17.24%	0.1455	0.1792	0.1724
MFCC + LSTM (Wazir et al. 2019)	6.89%	-0.0102	0.0472	0.0689
GFCC + CNN (Abdelmaksoud et al. 2021)	10.34%	0.0380	0.1043	0.1034
MFCC + CNN (Haque et al. 2020)	13.79%	0.0916	0.1757	0.1379

MFCC + CNN (Kherdekar and Naik, 2021)	0.68%	-0.0118	0.0153	0.0689
Hybrid CNN + BiLSTM (Passricha and Aggarwal, 2020)	0.68%	-0.0470	0.0309	0.0689
Proposed Method (*)	24.13%	0.1462	0.1802	0.2413

Table 6.12 Ablation studies of various NLP techniques in three Asian languages

Language	NLP Techniques	Training Accuracy	Test Accuracy
Arabic	Unigrams	100%	56.25%
	Bigrams	100%	59.38%
	Trigrams	100%	56.20%
Tamil	Unigrams	95.35%	62.79%
	Bigrams	100%	69.77%
	Trigrams	96.51%	61.36%
Vietnamese	Unigrams	90.01%	17.24%
	Bigrams	92.73%	27.59%
	Trigrams	92.73%	13.33%

6.3 SIGNIFICANT OUTCOMES

This research explores the classification of low-resource single-word audio datasets using a novel approach: a dense model formed by the late fusion of phoneme embeddings and bigram embeddings. What sets our approach apart is the application of late fusion to a single voice dataset with limited resources, a novel approach in itself. We found that the performance of the low-resource keyword spotting dataset significantly improved when integrating phoneme embeddings and bigrams embeddings into a 5-layered dense model with batch normalization. Our experiments utilized the MSWC dataset, featuring natural speaker audio recordings and ten-word categories. The results from our proposed approach for spoken word classification were promising. Interestingly, we observed that text transcripts can have a substantial impact on spoken word classification, outperforming audio-based features. We compared our approach with existing SOTA methods and conducted an ablation analysis of various NLP techniques for the selected Asian languages.

The next section In the same chapter discusses the crafting and Implementation of a classification framework for phonological and morphological features using pre-trained networks in spoken word recognition.

In this segment, we propose a supervised strategy for SWR in a resource-constrained environment within a ML dataset. Addressing a gap in the current SOTA methods, our approach integrates morphology and phonology for comprehending spoken text. The MSWC provides raw audio files in .OPUS format. To extract text transcripts, we employ the pre-trained Arabic Large xlsr-Wav2Vec2-53 transformer. Our experimental design consists of two stages, utilizing two forms of text transcripts: "buckwalter transliteration" and "Arabic script." In the initial stage, we convert the buckwalter transliteration form to phonemes using the CMU pronouncing dictionary and an Arabic-based grapheme 2-phoneme model. The obtained phonemes are then transformed into vectors through character n-gram-based subword embeddings from FastText. In the second stage, Arabic scripts are processed into stems using a Stemming algorithm, and the resulting stemmed Arabic script is further converted to unigrams. Transitioning from unigrams to vectors, we employ FastText word embeddings. To ensure consistency, we concatenate and pad the vectors in both scenarios. For the subsequent analysis, a three-layered dense and batch normalization model is employed, receiving the collected vectors to generate probabilistic scores. The outcomes of the two stages are averaged for result calculation. Comparing our findings with the SOTA approach, the results demonstrate a satisfactory performance, validating the effectiveness of our proposed methodology. This research contributes to the advancement of SWR, particularly in ML datasets, under resource constraints, and incorporates novel techniques involving morphology and phonology. This section makes significant contributions in the following areas:

- Extraction of Arabic script and buckwalter transliteration from text transcripts using transformers for the purpose of spoken word classification on Arabic Multilingual Spoken Words.

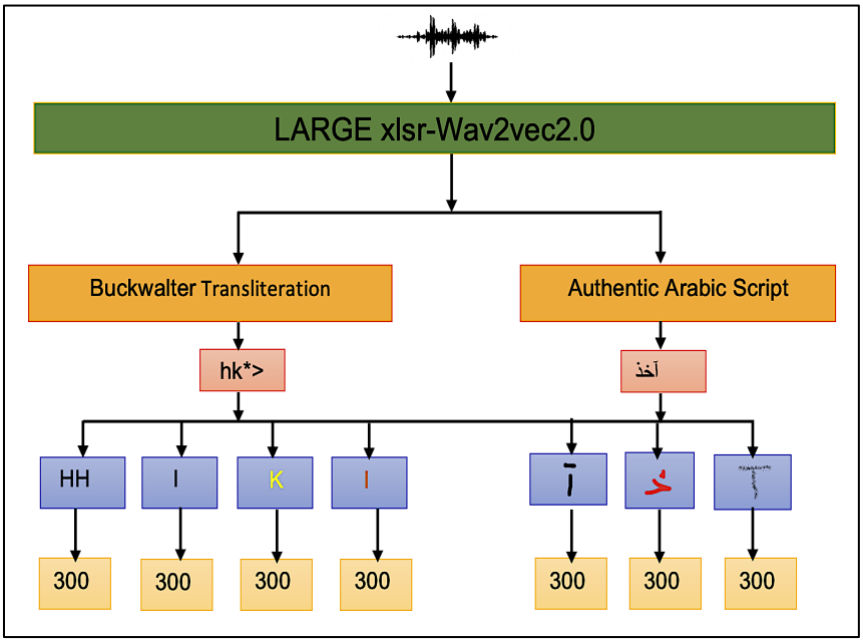
- Conversion of extracted buckwalter transliteration to phonemes using the G2P model, followed by transformation into embeddings using Fast-Text word embeddings. Concatenation and padding techniques are applied to ensure vector uniformity.
- Transformation of authentic Arabic Script into unigrams and subsequent conversion to embeddings using Fast-Text word embeddings.
- Utilization of the Dense + Batch Normalization model to process both sets of embeddings. The combination of unigrams and the phoneme technique is assessed using accuracy and f1-score parameters, demonstrating superior performance compared to the SOTA approach.

The remainder of this objective is organized as follows: Section 6.4 details the proposed modeling approach, encompassing early and late fusion techniques. Section 6.5 outlines the experimental design and presents a summary of the findings. Finally, Section 6.6 concludes the objective, suggesting avenues for future research.

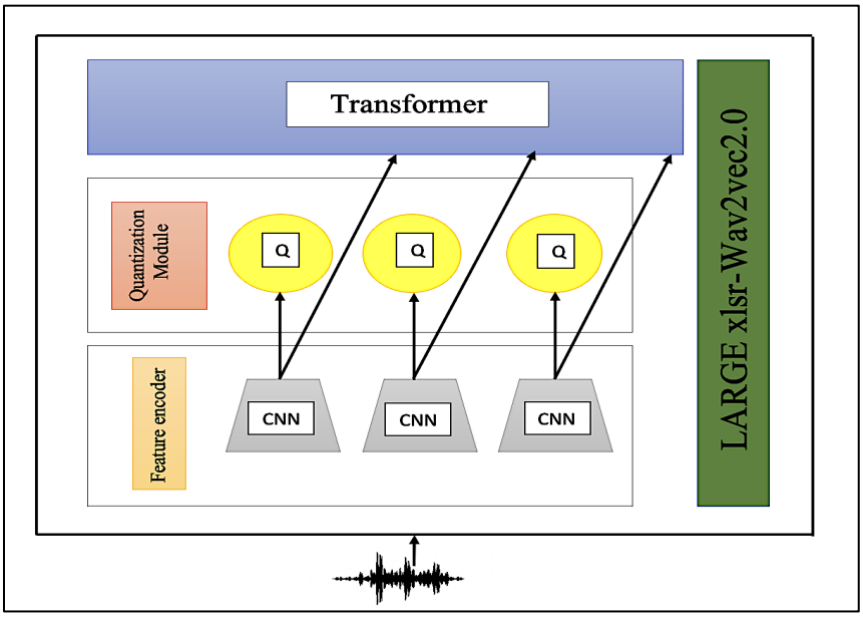
6.4 Proposed Approach

6.4.1 Speech-To-Text Transcript Extraction And Pre-Processing

In our proposed method, we converted raw audio to text using a pretrained Arabic LARGE xlsr-Wav2Vec2-53 model, known for its effectiveness in cross-lingual pretraining and supporting low-resource speech understanding. Combining CTC for ASR and transformers, as depicted in Figure 6.8, provides efficient pre-trained models across various frameworks and modalities. ML masked model pretraining, backed by wav2vec 2.0, and neural style transfer contribute to task versatility. Transcripts are obtained in both Buckwalter transliteration and Arabic script. The Buckwalter transliteration is further processed into phonemes using a G2P model and transformed into vectors with sub-word embeddings. Meanwhile, the Arabic script is initially transformed to its root form using the Arabic Porter Stemmer, and the stemmed script is further converted into unigrams using NLP techniques. In our approach, text transcripts closely approximate actual words, thanks to the use of a pre-trained model.



a)



b)

Figure 6.8. a) The generation of text transcripts and embeddings. b) LARGE xlsr-Wav2Vec2-53

6.4.2 Phoneme And Authentic Arabic Embeddings

The acquired phonemes are transformed into 300-dimensional vectors using FastText subword modeling, leveraging a pre-trained set of two million word vectors trained on Web Crawl data (600 billion tokens). These pre-trained word representations, derived from extensive text corpora encompassing news collections, Wikipedia, and web crawls, are

widely employed in various text-based applications (Mikolov et al. 2017). Simultaneously, the second text transcript, derived from authentic Arabic script using the pre-trained Arabic LARGE xlsr-Wav2Vec2-53 model, undergoes a distinct process. The authentic Arabic text is converted into root words, or "stems," utilizing the Porter Stemmer (Porter, 1980), chosen for its nuanced treatment of word dissections compared to other stemmer algorithms. The stems of the authentic Arabic text are further transformed into words-to-characters (unigrams), and subsequently converted into pre-trained word embeddings for the Arabic language. The standard Arabic word segmenter ICU tokenizer is employed in this context. FastText is utilized to train the unigrams for pre-trained word vectors using Common Crawl and Wikipedia. These models are trained employing CBOW with position weights in a 300-dimensional space, character n-grams of length 5, a window size of 5, and 10 negatives. Figure 6.8 illustrates the process of generating two forms of text transcripts from LARGE xlsr-Wav2Vec2-53. The buckwalter transliteration transcripts are converted to phonemes using the CMU pronouncing dictionary (Weide, 1998). These phonemes are further transformed into vectors, serving as input to the 3-layered dense + Batch Normalization (BN) model, from which probabilistic scores are extracted and stored. Conversely, the text transcript from authentic Arabic script is processed into characters or unigrams, and Arabic FastText Embeddings are applied in a 300-dimensional size per unigram. The resulting embeddings, concatenated and padded for uniformity, are then fed into the 3-layered dense model, yielding probabilistic scores that are extracted and stored.

Figure 6.8 showcases the text transcript generation process using the LARGE xlsr-Wav2Vec2-53 model. For Buckwalter transcripts, the CMU dictionary transforms them into phonemes, which are then converted into vectors. These vectors serve as input for a 3-layer dense + Batch Normalization (BN) model, generating scores. In the case of Arabic script, the LARGE xlsr-Wav2Vec2-53 model produces an authentic transcript. The script is further processed into unigrams, applied to Arabic FastText Embeddings (300-dimensional each). Concatenated and padded unigrams are fed into a 3-layer dense model, storing resultant scores. Figure 6.9 depicts the flow of information from both transcripts to the neural network, resulting in an average score that combines early and late fusion, creating a hybrid approach.

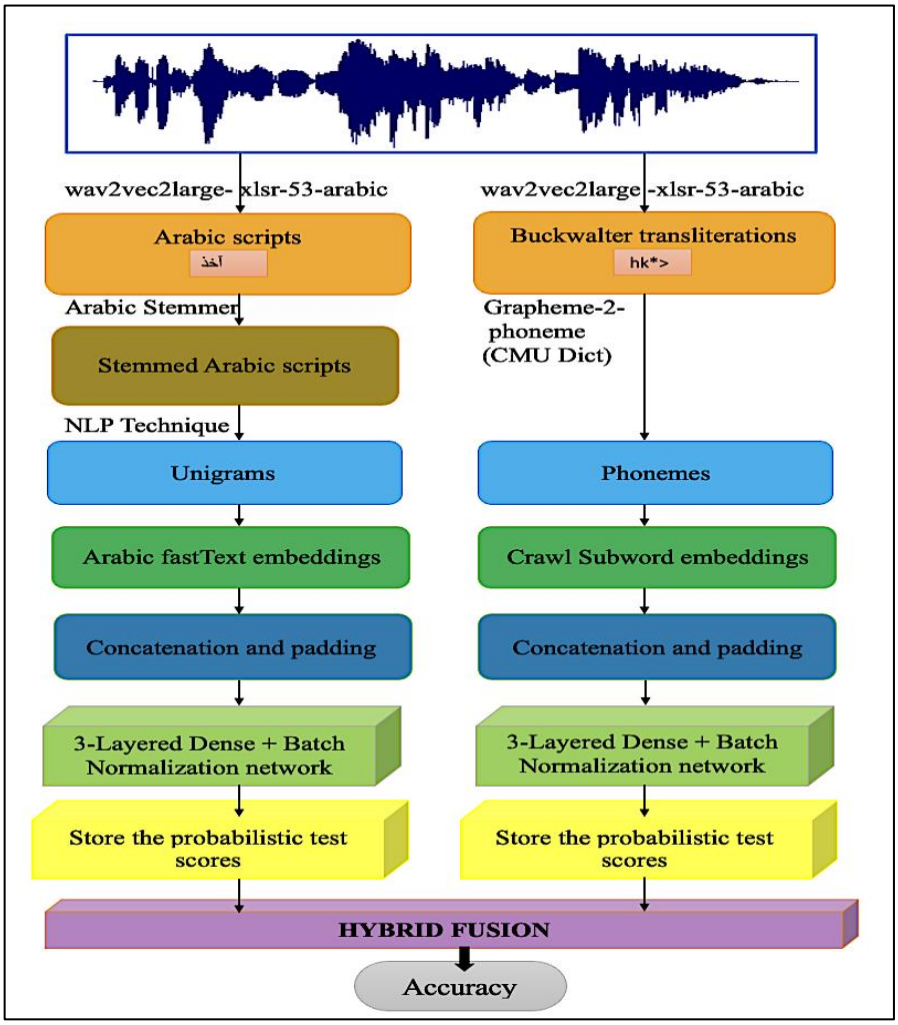


Figure 6.9. Proposed pipelines for Arabic spoken word recognition

6.5 Experiment And Discussions

6.5.1 Dataset

The MSWC, licensed under CC-BY 4.0, serves commercial keyword and spoken phrase search applications and academic research. With audio instances from over 5 billion people speaking 50 languages, the dataset contains 23.4 million 1-second spoken instances and 340,000 keywords. Applications range from contact centres to consumer electronics. The MSWC database includes spoken words in 50 languages, categorized by resource availability. However, for focused research on 10 and 50 spoken word categories, the dataset has been intentionally skewed, resulting in 94 and 722 samples, respectively. Despite its size, modifications cater to specific academic and commercial research objectives, emphasizing real-world applications. The MSWC was adapted for Arabic language categorization, aligning with the study's focus on academic research, commercial applications, keyword

spotting, and spoken term search. Despite the dataset's substantial size, it has been strategically modified for the specific exploration of 10- and 50-word categories, resulting in a tailored dataset for targeted research objectives.

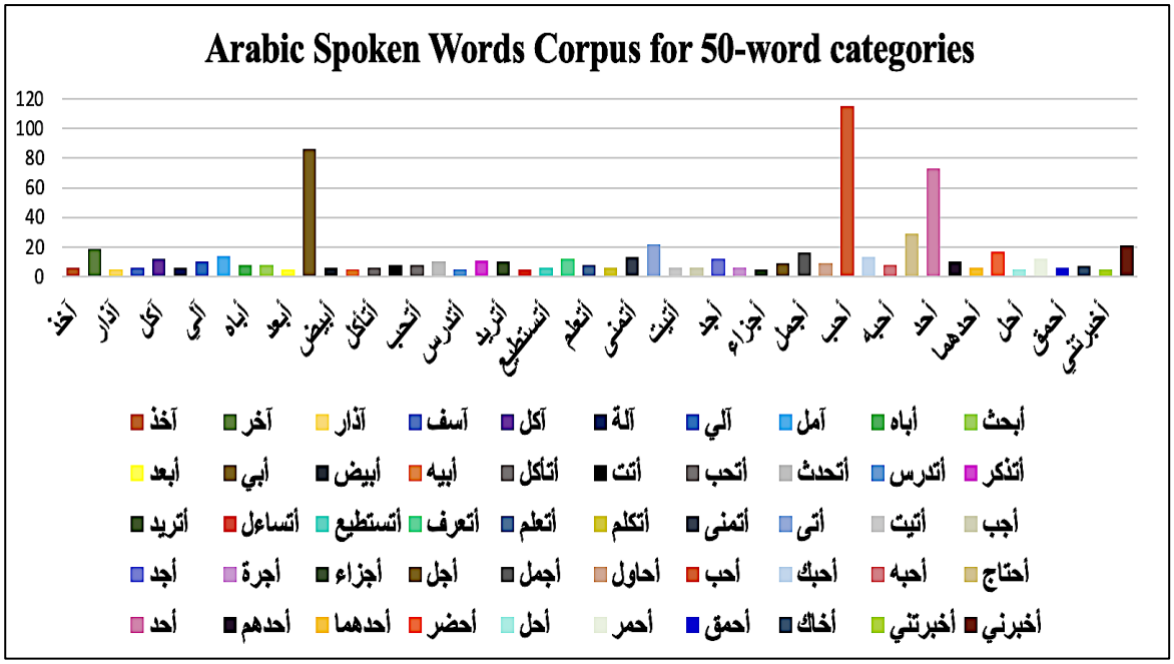


Figure 6.10. The Number of samples for 50-spoken word categories

6.5.2 Experimental Setup

The investigations were conducted using Python 3.10.4 on a MAC OS High Sierra system equipped with an Intel Core i5 CPU featuring Intel Graphics, operating at a clock speed of 1.8 GHz. The experimental setup employed a 70:30 train:test split ratio with 5-fold cross-validation. The optimization process utilized the Adam optimizer, employing a batch size of 16, 100 epochs, and a learning rate of 0.01. Nonlinearity was introduced into the model through the use of the "ReLU" activation function.

6.5.3 Results And Discussions

Utilizing an under-resourced and highly unbalanced dataset, Arabic SWR is approached using the pre-trained Arabic LARGE xlsr-Wav2Vec2-53 model. Text transcripts are extracted in buckwalter transliteration and authentic Arabic script, with the former

transformed into phonemes and further into 300-dimensional vectors using FastText subword modeling.

Experimentation targets 10 and 50-word categories, posing challenges in handling dataset imbalance. The authentic Arabic script is processed into stems and unigrams, converted into vectors via concatenation and padding. Vector dimensions are 3900 X 1 for both 10 and 50-spoken word categories. Buckwalter transliteration text transcripts undergo phoneme conversion, resulting in vectors of sizes 2700 X 1 and 3000 X 1 for 10 and 50-spoken word categories, respectively. Hyperparameters, define a 3-layered model with batch normalization, "ReLU" activation, and Adam optimizer. A 70:30 train:test split ratio with 5-fold cross-validation is applied, using a batch size of 16, 100 epochs, and a learning rate of 0.01. The dataset's imbalance is addressed through concatenation and padding, maintaining consistency in vector dimensions for both transcript forms. Early feature fusion is achieved by training independently on the padded features for 10 and 50 spoken words. In comparison, the buckwalter transliteration form surpasses the Arabic script in accuracy. Stemming is applied to the latter, but resource constraints limit its study. Dataset imbalance further complicates understanding spoken Arabic speech. For the 10-spoken words category in buckwalter transliteration, after concatenation and padding, the output shape is (2700) with a total parameter count of 1,534,410. For the 50-spoken word category, the output shape is (3000), with a total parameter count of 1,690,610, comprising 1,689,074 trainable and 1,536 non-trainable parameters. The input shape of Arabic script padded vectors is (3900) for both 10 and 50 spoken words. These vectors, after training, are fed into a flattened layer and three dense layers, totalling 2,148,810 parameters. The early fusion accuracy for phoneme and unigram embeddings in the MSWC is 68.96% and 44.83%, respectively, for 10-word categories and 67.28% and 52.78% for 50-word categories. Late fusion, combining the probabilistic scores from both procedures, yields 72.41% and 70.97% accuracy for 10 and 50 spoken word categories, respectively.

The confusion matrix in Figures 6.11 and 6.12 shows the early fusion hypothesis results, highlighting the highest accuracy for spoken word-2 category for 10 classes and category-37 for 50 spoken word categories. Limited dataset quantity impacts model performance, with varying accented speech contributing to lower accuracies. Comparison with established experiments reveals our method's superiority over MFCC-based models, achieving higher

accuracy than CNN-based and hybrid CNN + BiLSTM models. Evaluation measures for specificity, FNR, and FDR are high across all classes, while sensitivity and PPV show variations. The utilization of linguistic features, particularly bigrams and phonemes, in transformer-generated text transcripts contributes to the success of our approach in handling low-resourced languages and skewed datasets. Table 6.13 presents the classified spoken words for both 10- and 50-word categories following the late fusion approach. The outcomes of late fusion demonstrate notable satisfaction when compared with the methods outlined in Table 6.15. The utilization of text-based features, as opposed to audio-based features, contributes to the superior quality of the final results. In Figure 6.13, the classification report illustrates the performance of our proposed method across 10 Arabic spoken word categories. The evident superiority of our approach, characterized by substantial accuracy and minimal WER, surpasses that of the comparative methods.

Table 6.13 Number of classified spoken words after training for 50-word categories

Spoken word categories	Number of classified spoken words	Accuracy	WER
آخذ	1	1.00000000	0.00
آخر	4	0.99078341	0.01
آذار	0	0.99539171	0.00
آسف	2	0.99539171	0.00
آكل	2	0.98617512	0.01
آلة	0	0.98617512	0.01
آلي	4	0.99078341	0.01
آمل	3	0.96313364	0.04
أباه	0	0.97235023	0.03
أبحث	2	1.00000000	0.00
أبعد	0	0.99539171	0.00
أبي	22	0.97235023	0.03
أبيض	0	0.99539171	0.00
أبيه	0	0.99539171	0.00
أتأكل	1	0.99078341	0.00
أتت	1	0.98617512	0.01
أتحب	3	0.98617512	0.01
أتحدث	2	0.99078341	0.00
أتدرس	1	0.99539171	0.00
أتذكر	2	0.99078341	0.00
أتريد	2	0.99078341	0.00
أتساءل	0	0.99539171	0.00
أتستطيع	2	0.99078341	0.01
أتعرف	1	0.99078341	0.01
أتعلم	1	0.99539171	0.00
أتكلم	1	0.98617512	0.01
أتمنى	2	0.99078341	0.00
أتى	3	0.94930876	0.05
أتيت	2	0.99078341	0.01
أجب	3	0.99539171	0.00
أجد	2	0.98617512	0.01

أجرة	0	0.99539171	0.00
أجزاء	1	0.99539171	0.00
أجل	1	0.99078341	0.01
أجمل	4	0.99539171	0.00
أحاول	2	0.99539171	0.00
أحب	31	0.94930876	0.05
أحبك	1	0.98617512	0.01
أحبه	0	0.99078341	0.01
أحتاج	10	0.98156682	0.02
أحد	16	0.94470046	0.06
أحدهم	0	0.98617512	0.01
أحدهما	0	0.99078341	0.01
أحضر	5	0.99539171	0.00
أحل	1	0.99078341	0.00
أحمر	4	0.99078341	0.00
أحمق	1	0.99078341	0.00
أخاك	0	0.99078341	0.00
أخبرتني	1	1. 00000000	0.00
أخبرني	5	1. 00000000	0.00

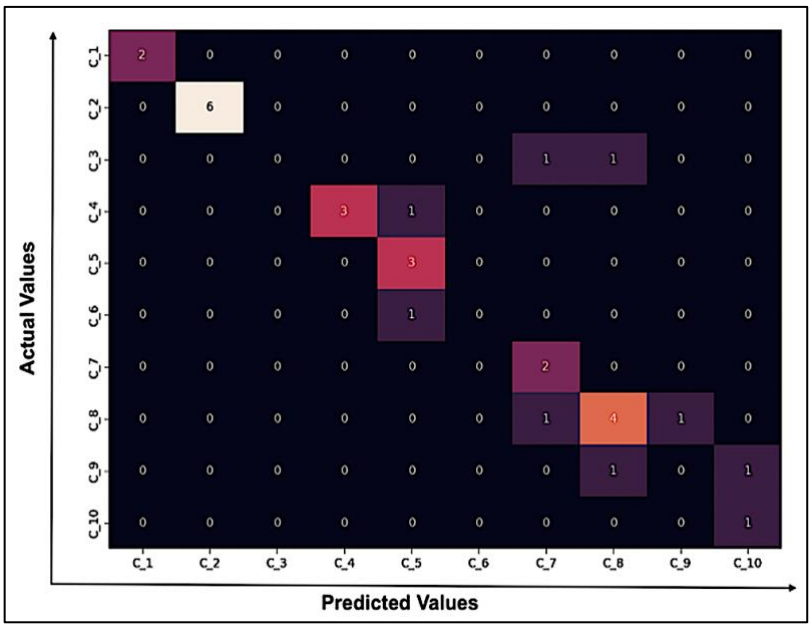


Figure 6.11. Confusion matrix for 10-spoken word categories after hybrid fusion

Table 6.14 Proposed evaluation metric with 5-fold cross-validation

Accuracy		Mathew's correlation coefficient (MCC)	Macro	Micro
Proposed Method	70.97%	0.6166	0.4971	0.7097

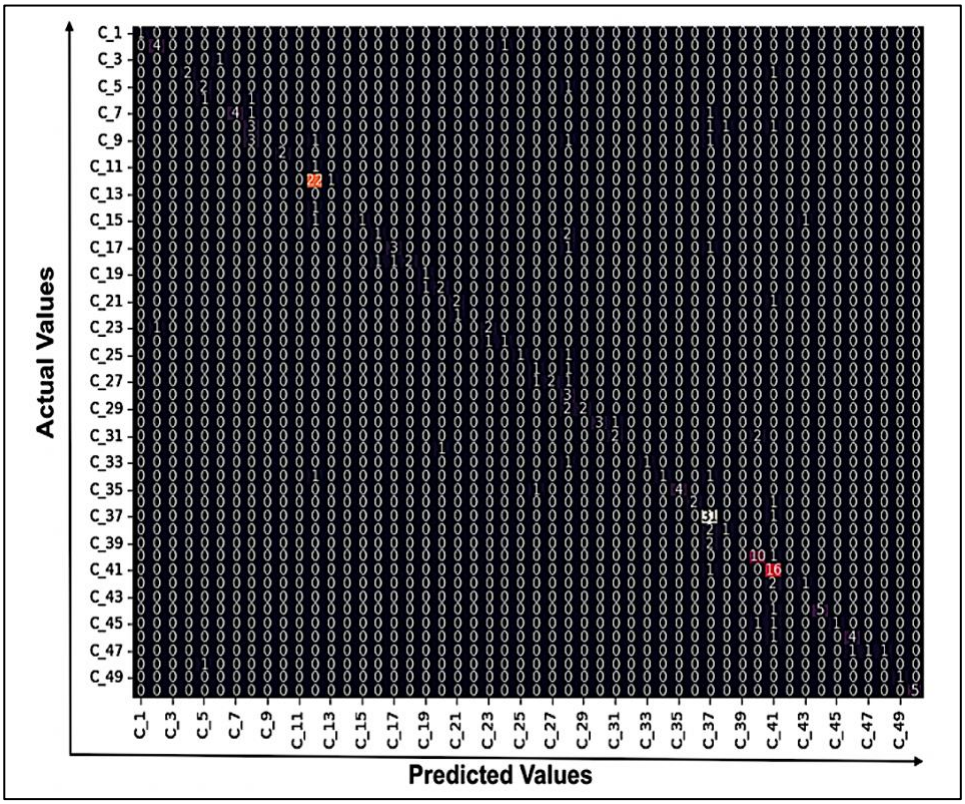


Figure 6.12. Confusion matrix for 50-spoken word categories after hybrid fusion

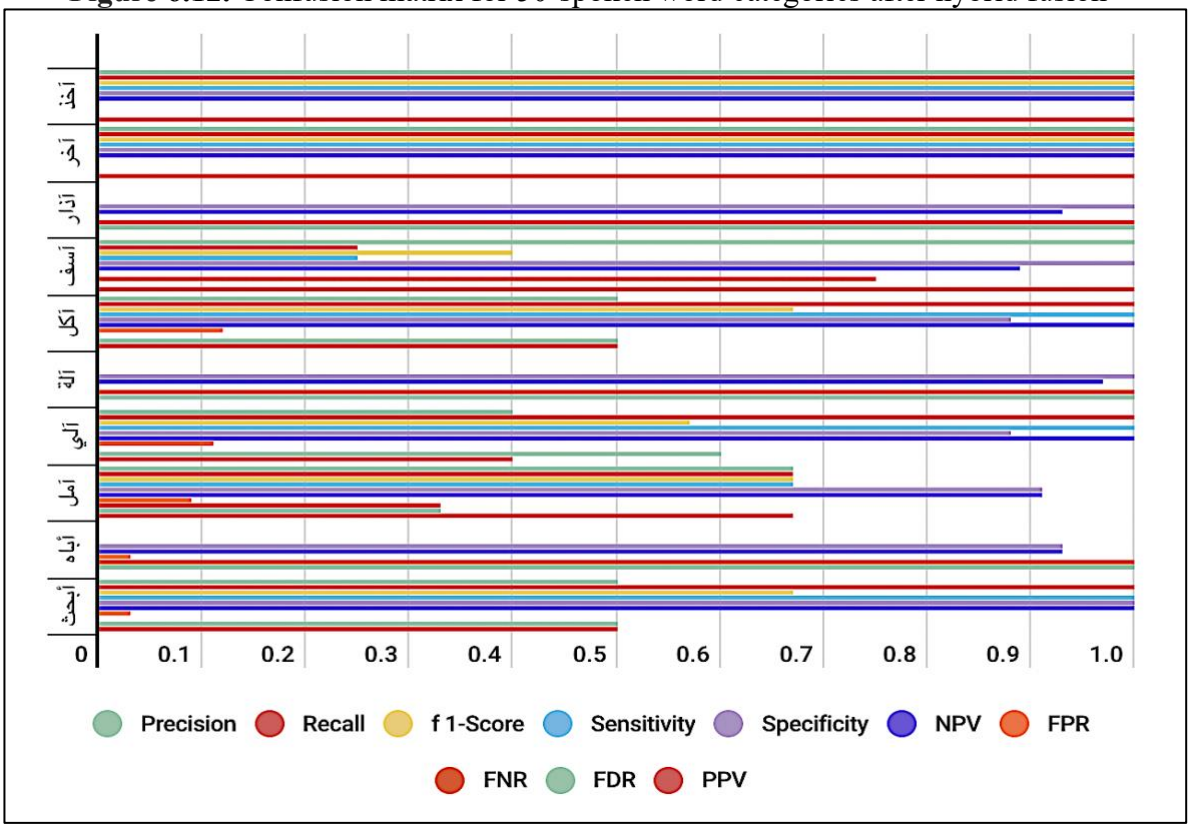


Figure 6.13. Classification results of late fusion of Buckwalter transliteration and authentic Arabic Script to detect Arabic Spoken Words for 10-word categories

Table 6.15 Accuracy results on comparison and proposed methods with 5-fold cross-validation

Spoken word categories	Precision	Recall	f 1-Score	Sensitivity	Specificity	NPV	FPR	FNR	FDR	PPV
آخذ	0.00 (1) 0.25 (2) 0.25 (3) 0.20 (4) 0.00 (5) 1.00 (*)	0.00 (1) 0.10 (2) 0.10 (3) 0.50 (4) 0.00 (5) 1.00 (*)	0.00 (1) 0.14 (2) 0.14 (3) 0.29 (4) 0.00 (5) 1.00 (*)	0.00 (1) 0.10 (2) 0.10 (3) 0.50 (4) 0.00 (5) 1.00 (*)	1.00 (1) 0.95 (2) 0.95 (3) 0.85 (4) 0.92 (5) 1.00 (*)	0.86 (1) 0.85 (2) 0.85 (3) 0.96 (4) 0.88 (5) 1.00 (*)	0.00 (1) 0.05 (2) 0.05 (3) 0.15 (4) 0.08 (5) 0.00 (*)	1.00 (1) 0.90 (2) 0.90 (3) 0.50 (4) 1.00 (5) 0.00 (*)	1.00 (1) 0.75 (2) 0.75 (3) 0.80 (4) 1.00 (5) 0.00 (*)	0.00 (1) 0.25 (2) 0.25 (3) 0.20 (4) 0.00 (5) 1.00 (*)
آخر	0.00 (1) 1.00 (2) 0.67 (3) 0.00 (4) 0.00 (5) 1.00 (*)	0.00 (1) 0.50 (2) 0.50 (3) 0.00 (4) 0.00 (5) 1.00 (*)	0.00 (1) 0.67 (2) 0.57 (3) 0.00 (4) 0.00 (5) 1.00 (*)	0.00 (1) 0.50 (2) 0.50 (3) 0.00 (4) 0.00 (5) 1.00 (*)	1.00 (1) 1.00 (2) 0.98 (3) 1.00 (4) 0.97 (5) 1.00 (*)	0.94 (1) 0.97 (2) 0.97 (3) 0.79 (4) 0.94 (5) 1.00 (*)	0.00 (1) 0.00 (2) 0.02 (3) 0.00 (4) 0.03 (5) 0.00 (*)	1.00 (1) 0.50 (2) 0.50 (3) 1.00 (4) 1.00 (5) 0.00 (*)	1.00 (1) 0.00 (2) 0.33 (3) 1.00 (4) 1.00 (5) 0.00 (*)	0.00 (1) 1.00 (2) 0.67 (3) 0.00 (4) 0.00 (5) 1.00 (*)
آذار	0.18 (1) 0.17 (2) 0.17 (3) 0.00 (4) 0.10 (5) 0.00 (*)	0.33 (1) 0.29 (2) 0.29 (3) 0.00 (4) 0.20 (5) 0.00 (*)	0.24 (1) 0.21 (2) 0.21 (3) 0.00 (4) 0.13 (5) 0.00 (*)	0.33 (1) 0.28 (2) 0.28 (3) 0.00 (4) 0.20 (5) 0.00 (*)	0.67 (1) 0.83 (2) 0.83 (3) 0.89 (4) 0.66 (5) 1.00 (*)	0.82 (1) 0.91 (2) 0.91 (3) 0.92 (4) 0.82 (5) 0.93 (*)	0.33 (1) 0.17 (2) 0.17 (3) 0.11 (4) 0.34 (5) 0.00 (*)	0.67 (1) 0.71 (2) 0.71 (3) 1.00 (4) 0.80 (5) 1.00 (*)	0.82 (1) 0.83 (2) 0.83 (3) 1.00 (4) 0.91 (5) 1.00 (*)	0.18 (1) 0.17 (2) 0.17 (3) 0.00 (4) 0.10 (5) 0.00 (*)
آسف	0.14 (1) 0.60 (2) 0.60 (3) 0.00 (4) 0.00 (5) 1.00 (*)	0.80 (1) 0.50 (2) 0.50 (3) 0.00 (4) 0.00 (5) 0.25 (*)	0.24 (1) 0.55 (2) 0.55 (3) 0.00 (4) 0.00 (5) 0.40 (*)	0.80 (1) 0.50 (2) 0.50 (3) 0.00 (4) 0.00 (5) 0.25 (*)	0.59 (1) 0.97 (2) 0.97 (3) 1.00 (4) 0.98 (5) 1.00 (*)	0.97 (1) 0.95 (2) 0.95 (3) 0.86 (4) 0.89 (5) 0.89 (*)	0.41 (1) 0.03 (2) 0.03 (3) 0.00 (4) 0.02 (5) 0.00 (*)	0.20 (1) 0.50 (2) 0.50 (3) 1.00 (4) 1.00 (5) 0.75 (*)	0.87 (1) 0.40 (2) 0.40 (3) 1.00 (4) 1.00 (5) 0.00 (*)	0.14 (1) 0.60 (2) 0.60 (3) 0.00 (4) 0.10 (5) 1.00 (*)
آكل	0.00 (1) 0.00 (2) 1.00 (3) 0.00 (4) 0.00 (5) 0.50 (*)	0.00 (1) 0.00 (2) 0.20 (3) 0.00 (4) 0.00 (5) 1.00 (*)	0.00 (1) 0.00 (2) 0.33 (3) 0.00 (4) 0.00 (5) 0.67 (*)	0.00 (1) 0.00 (2) 0.20 (3) 0.00 (4) 0.00 (5) 1.00 (*)	0.90 (1) 1.00 (2) 1.00 (3) 1.00 (4) 0.94 (5) 0.88 (*)	0.91 (1) 0.92 (2) 0.94 (3) 0.90 (4) 0.95 (5) 1.00 (*)	0.10 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.06 (5) 0.12 (*)	1.00 (1) 1.00 (2) 0.80 (3) 1.00 (4) 1.00 (5) 0.00 (*)	1.00 (1) 1.00 (2) 0.00 (3) 1.00 (4) 1.00 (5) 0.50 (*)	0.00 (1) 0.00 (2) 1.00 (3) 0.00 (4) 0.00 (5) 0.05 (*)
آلة	0.00 (1) 1.00 (2) 1.00 (3) 0.00 (4) 0.00 (5) 0.00 (*)	0.00 (1) 0.33 (2) 0.33 (3) 0.00 (4) 0.00 (5) 0.00 (*)	0.00 (1) 0.50 (2) 0.50 (3) 0.00 (4) 0.00 (5) 0.00 (*)	0.00 (1) 0.33 (2) 0.33 (3) 0.00 (4) 0.00 (5) 0.00 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 0.93 (5) 1.00 (*)	0.94 (1) 0.97 (2) 0.97 (3) 0.96 (4) 0.92 (5) 0.97 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.06 (5) 0.00 (*)	1.00 (1) 0.67 (2) 0.67 (3) 1.00 (4) 1.00 (5) 1.00 (*)	1.00 (1) 1.00 (2) 0.00 (3) 1.00 (4) 1.00 (5) 1.00 (*)	0.00 (1) 1.00 (2) 1.00 (3) 0.00 (4) 0.00 (5) 0.00 (*)
آلي	0.00 (1) 0.22 (2) 0.22 (3) 0.50 (4) 0.00 (5) 0.40 (*)	0.00 (1) 0.33 (2) 0.33 (3) 0.50 (4) 0.00 (5) 1.00 (*)	0.00 (1) 0.27 (2) 0.27 (3) 0.50 (4) 0.00 (5) 0.57 (*)	0.00 (1) 0.33 (2) 0.33 (3) 0.50 (4) 0.00 (5) 1.00 (*)	0.97 (1) 0.88 (2) 0.88 (3) 0.96 (4) 1.00 (5) 0.88 (*)	0.92 (1) 0.93 (2) 0.93 (3) 0.96 (4) 0.94 (5) 1.00 (*)	0.03 (1) 0.11 (2) 0.12 (3) 0.04 (4) 0.00 (5) 0.11 (*)	1.00 (1) 0.67 (2) 0.67 (3) 0.50 (4) 1.00 (5) 0.00 (*)	1.00 (1) 0.78 (2) 0.78 (3) 0.50 (4) 1.00 (5) 0.60 (*)	0.00 (1) 0.22 (2) 0.22 (3) 0.50 (4) 0.00 (5) 0.04 (*)
آمل	0.00 (1) 0.19 (2) 0.20 (3) 0.33 (4) 0.21 (5) 0.67 (*)	0.00 (1) 0.38 (2) 0.38 (3) 0.67 (4) 0.46 (5) 0.67 (*)	0.00 (1) 0.25 (2) 0.26 (3) 0.44 (4) 0.29 (5) 0.67 (*)	0.00 (1) 0.38 (2) 0.38 (3) 0.67 (4) 0.00 (5) 0.67 (*)	1.00 (1) 0.58 (2) 0.62 (3) 0.65 (4) 0.56 (5) 0.91 (*)	0.80 (1) 0.79 (2) 0.80 (3) 0.88 (4) 0.81 (5) 0.91 (*)	0.00 (1) 0.41 (2) 0.38 (3) 0.35 (4) 0.43 (5) 0.09 (*)	1.00 (1) 0.62 (2) 0.62 (3) 0.33 (4) 0.54 (5) 0.33(*)	1.00 (1) 0.81 (2) 0.80 (3) 0.67 (4) 0.54 (5) 0.33(*)	0.00 (1) 0.19 (2) 0.20 (3) 0.33 (4) 0.21 (5) 0.67 (*)
أباه	0.00 (1) 0.33 (2) 0.33 (3) 0.17 (4) 0.00 (5) 0.00 (*)	0.00 (1) 0.25 (2) 0.25 (3) 0.50 (4) 0.00 (5) 0.00 (*)	0.00 (1) 0.29 (2) 0.29 (3) 0.25 (4) 0.00 (5) 0.00 (*)	0.00 (1) 0.25 (2) 0.25 (3) 0.50 (4) 0.46 (5) 0.00 (*)	0.93 (1) 0.93 (2) 0.93 (3) 0.81 (4) 1.00 (5) 0.96 (*)	0.92 (1) 0.90 (2) 0.90 (3) 0.96 (4) 0.88 (5) 0.93 (*)	0.06 (1) 0.07 (2) 0.07 (3) 0.19 (4) 0.00 (5) 0.03 (*)	1.00 (1) 0.75 (2) 0.75 (3) 0.50 (4) 1.00 (5) 1.00 (*)	1.00 (1) 0.67 (2) 0.67 (3) 0.83 (4) 1.00 (5) 1.00 (*)	0.00 (1) 0.33 (2) 0.33 (3) 0.17 (4) 0.00 (5) 0.00 (*)
أبحث	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.50 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 1.00 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.67 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 1.00 (*)	0.95 (1) 1.00 (2) 1.00 (3) 0.96 (4) 1.00 (5) 0.96 (*)	0.94 (1) 0.94 (2) 0.94 (3) 0.96 (4) 0.92 (5) 1.00 (*)	0.05 (1) 0.00 (2) 0.00 (3) 0.03 (4) 0.00 (5) 0.03 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 0.00 (*)	1.00 (1) 1.00 (2) 1.00 (3) 1.00 (4) 1.00 (5) 0.50 (*)	0.00 (1) 0.00 (2) 0.00 (3) 0.00 (4) 0.00 (5) 0.50 (*)

6.6 SIGNIFICANT OUTCOMES

In this study, we propose a supervised technique for SWR in a resource-limited ML corpus, addressing the underexplored aspects of morphology and phonology. Using early fusion, combining phone embeddings and unigrams yields accuracies of 68.96% and 44.83% for 10 spoken word categories; for 50 categories, accuracies are 67.28% and 52.78%. Late fusion of probabilistic scores from both approaches outperforms prior methods, achieving accuracies of 72.41% and 70.97% for 10 and 50 spoken word categories. Our findings highlight the significant impact of phonemes from raw audio on multilingual SWR, with future work aiming to expand to more languages and explore the role of audio phonemes.

In the next chapter, we have investigated the impact of a fusion framework on recognizing spoken words, incorporating both raw audio and speech transcriptions.

CHAPTER 7

DESIGN AND DEVELOPMENT OF FUSION FRAMEWORK FOR SPOKEN WORD RECOGNITION FROM RAW-AUDIO AND SPEECH TRANSCRIPTIONS

In this chapter, our research focuses on the effective integration of multi-modal data, encompassing imperfect text transcripts and raw audio, within a DL framework for automatic speech recognition. Our approach involves the late fusion of audio and text modalities during the recognition process. We introduce a novel model, termed the SA-deep BiLSTM, which is designed to independently process audio and text data. For training each feature type, we employ the SA-deep BiLSTM model, comprising five BiLSTM layers with a self-attention module positioned between the third and fourth layers. This model accommodates both linguistic data, such as word stems extracted from text transcripts, and acoustic features like MFCC and Mel-spectrograms. To vectorize linguistic data, we employ GloVe word embeddings. By fusing the posterior class probabilities obtained from SA-deep BiLSTM models trained on individual modalities, we achieve a remarkable accuracy of 98.80% when classifying 10-word categories within the Google speech command dataset.
[4]

[4] The contents of this chapter are submitted/accepted/under review in:
"Deep fusion framework for speech command recognition using acoustic and linguistic features - Multimedia Tools and Applications (2023): Vol 82." 38667–38691 (2023), https://doi.org/10.1007/s11042-023-15118-1. (IF: 3.6).
&
"Multimodal integration of mel spectrograms and text transcripts for enhanced automatic speech recognition: Leveraging extractive transformer-based approaches and late fusion strategies" Currently under review in Computational Intelligence (2024)". (IF: 2.1).

Our approach undergoes rigorous testing using the Google speech command dataset, including ablation analysis, which demonstrates its superiority over SOTA methods. This performance improvement is attributed to the high classification accuracies we consistently attain. In summary, our research presents an innovative SA-deep BiLSTM model that excels in combining multi-modal data for automatic speech recognition, and it achieves outstanding results, particularly evident in the high classification accuracies achieved when applied to challenging datasets.

Our work contributes significantly in the following ways:

1. We introduce a SA-deep BiLSTM architecture that autonomously learns acoustic and linguistic features from raw audio and text transcripts, respectively.
2. Linguistic features, represented by word stems in the text transcripts, are vectorized using GloVe embeddings, while acoustic features include MFCC and Mel-spectrogram.
3. We employ late fusion on the probabilistic predictions generated by individual SA-deep BiLSTM models to identify speech commands.
4. To highlight the effectiveness of our fusion framework compared to models trained on smaller subsets of audio and text modalities, we present an ablation analysis.
5. Our work includes an extensive comparative analysis with SOTA methods, demonstrating the efficacy of our approach in speech command recognition.

The chapter is organized as follows: In Section 7.1, we introduce our advanced fusion framework, which combines acoustic and linguistic features in a DL context. Section 7.2 delineates our experimental setup, providing a concise summary of our methodology and the outcomes we obtained. Finally, in Section 7.3, we draw significant outcomes and offer insights into potential avenues for future research.

7.1 Proposed Deep Fusion Framework For Acoustic And Linguistic Features

In this study, we investigate the fusion of audio and text modalities for speech command recognition. We introduce a SA-deep BiLSTM model designed to classify speech commands. This model leverages acoustic features like MFCC and Mel-spectrogram, as well as linguistic features, specifically the word stems extracted from text transcripts. Our core hypothesis centres around the idea that the performance of the speech recognition system can be significantly enhanced by merging the decision streams from acoustic and linguistic modalities within a DL framework. To achieve this, we employ a late fusion strategy, which probabilistically combines the output predictions. The complete algorithm for our proposed approach is outlined below. We begin by describing the architecture of the proposed SA-deep BiLSTM model, which is employed to autonomously learn acoustic and linguistic modalities. The fundamental challenge with basic RNNs is their inability to effectively handle long sequences and capture critical information. To address this limitation and capture relevant information from extended data sequences, LSTMs were developed. LSTMs consist of a cell state (network's memory), a hidden state, and three crucial gates that facilitate the continuous flow of gradients, crucial for making predictions. The trio comprises an output gate, an input gate, and a forget gate. The forget gate plays a pivotal role in determining which information should be retained or discarded from the cell state. When data (x_t) and the previous hidden state (h_{t-1}) are provided as inputs, this gate, effectively modelled as a sigmoid function, outputs a value between 0 and 1 for each component of the cell state. A value of 0 signifies the discarding of information, while a value of 1 signifies retaining the information. The data (x_t) and the previous hidden state (h_{t-1}) first pass through a sigmoid layer in the output gate. The new hidden state (h_t) is then generated by multiplying the result with c_t (after passing through a "*tanh"* layer). The term "bi-directional LSTM" indicates the technique of first computing hidden states from front to rear and then in the reverse direction, followed by the combination of the two results. LSTMs accept input in the form of *samples* × *features* × *time-steps* and link each time-step input recursively to the previous memory. They are widely used for classifying sequential data, including audio and text. Let's denote x_t as the LSTM input, representing audio or text vectors from sequential data. ht represents the hidden state at the current timestamp, and h_{t-1} is the hidden state from the previous timestamp. The input and the previous hidden state combine to create the vector

X, serving as the input for the forget, input, and output gates. Weight matrices W_i, W_f, W_o, and biases b_i, b_f, b_o are determined during the training phase. σ represents the sigmoid function. In summary, the SA-deep BiLSTM model, with its LSTM architecture, plays a crucial role in processing sequential data, making it a valuable tool for tasks involving audio and text data analysis.

$$X = [h_{t-1}, x_t] \tag{7.1}$$

$$ft = \sigma\,(W_f\, X + b_f) \tag{7.2}$$

$$it = \sigma\,(W_i\, X + b_i) \tag{7.3}$$

$$o_t = \sigma\,(W_o\, X + b_o) \tag{7.4}$$

In (7.2) to (7.4), the symbols *f*, *i*, and *o* correspond to the gate activations. The "*tanh*" denotes the hyperbolic tangent function, and the "*" symbol indicates element-wise multiplication. In (7.5), c_{t-1} and c_t represent the previous and current cell states, respectively.

$$c_t = f_t * c_{t-1} + i_t * \tanh\,(W_c * X + b_c) \tag{7.5}$$

The hidden state at time-step t is computed as:

$$h_t = o_t * \tanh\,(c_t) \tag{7.6}$$

The primary distinction among LSTM, BiLSTM, and Deep BiLSTM lies in their directional processing. LSTM is unidirectional and retains information solely from the past. In contrast, BiLSTM processes inputs both from the past to the future and from the future to the past, effectively capturing information bidirectionally. Deep BiLSTM takes this bidirectional processing further by incorporating multiple recurrent layers, enhancing the overall efficiency of the BiLSTM model. In a BiLSTM, there exist two hidden layers, one for the forward pass and one for the backward pass. The final hidden state results from concatenating the hidden states computed in both the forward and backward passes, as demonstrated in (7.7).

$$h_t = [\overrightarrow{ht}\,;\,\overleftarrow{ht}] \tag{7.7}$$

Attention mechanisms play a crucial role in enabling the model to access contextual information by examining the neighbouring words of the target word. Self-attention, also

referred to as intra-attention, differs from the inter-attention mechanism proposed by Bahdanau et al. (2014). Attention mechanisms empower the model to directly observe and extract information from previous vectors. The practical value of attention mechanisms was demonstrated in the context of neural machine translation (NMT) (Vaswani et al., 2017). The computations for self-attention are outlined in (7.8) to (7.10).

$$l_t = \sum_{t'} \alpha_{t,t'} x_{t'} \tag{7.8}$$

$$\alpha_{t,t'} = \text{softmax}\,(\sigma(W_a h_{t,t'} + b_a)) \tag{7.9}$$

$$h_{t,t'} = \tanh\,(x_t^T W_t + x_{t'}^T W_x + b_t) \tag{7.10}$$

Our hypothesis posits that the incorporation of self-attention into the deep BiLSTM model would yield superior results when fusing information from audio and linguistics compared to learning from these modalities independently. We derive predictions from individual deep models trained on acoustic and linguistic features and combine these predictions using a soft fusion technique. Soft fusion of prediction probabilities is a well-established technique in machine learning, commonly used for aggregating predictions from individual models in an ensemble. In our work, we adopt a late fusion strategy, where we fuse the probabilistic predictions by either averaging or selecting the maximum probabilistic score associated with each class. The soft fusion process in our deep fusion framework is as follows: Let pc denote the posterior class probability associated with the spoken word category c. We combine the three probabilistic decision scores obtained from the SA-Deep BiLSTM model trained on MFCC, Mel-spectrogram (ms), and stem, using both maximum and average functions, as outlined below.

$$pc = \max\,(pc\,(mfcc), pc\,(ms), pc\,(stem)) \tag{7.11}$$

$$pc = \text{mean}\,(pc\,(mfcc), pc\,(ms), pc\,(stem)) \tag{7.12}$$

The class of the test sample is calculated:

$$class = \forall c\ argmax(pc) \tag{7.13}$$

To capture both the spectral characteristics of audio and the linguistic properties of the transcript, we process the data as follows:

1. We extract acoustic features, specifically MFCC and Mel-Spectrograms, from the raw .WAV audio files. These features are transformed into matrices with dimensions samples × features.
2. To preserve linguistic information, we employ the Porter stemming algorithm (Porter, 1999) on the text transcript obtained through the Google API. The stemmed transcript is then converted into a feature matrix using GloVe word embeddings (Pennington, 2014). In Figure 7.1, we depict the architecture of our proposed SA-deep BiLSTM model. The "MATRIX" in Figure 7.1 represents the feature vectors extracted from two modalities: audio and text. The frame length is standardized to 44 for all audio files. The dimensions of the MFCC matrix are (44 × 39), the Mel-spectrogram has a shape of (44 × 128), and the GloVe word-embedded vectors for the stems have a shape of (50 × 1).
3. After extracting the acoustic and linguistic features (MFCC, Mel-spectrogram, stem), each feature serves as input to the SA-deep BiLSTM model. The trained model generates posterior class probabilities for each test sample. This comprehensive approach allows us to capture the complex interplay of acoustic and linguistic information in the data.

Our deep BiLSTM model is structured with three initial BiLSTM layers, comprising 512, 256, and 128 units, followed by a self-attention layer with 128 units. Additionally, we incorporate two high-level BiLSTM layers with 256 and 128 units. Further, our model includes a dense layer with 32 units, a dropout layer with 32 units, and another dense layer with 10 units. In total, the model consists of five BiLSTM layers, rendering it considerably deep. The input features are processed through both forward and backward LSTMs to obtain the forward and backward hidden states. The final layer of our SA-deep BiLSTM model is a dense layer with a "softmax" activation function, responsible for generating class probability predictions. In Figure 7.1, L represents the cells of the forward LSTM, while L' represents the cells of the backward LSTM, forming one layer of the BiLSTM. Figure 7.2 provides a flowchart depicting how predictions are generated by the three deep models within our fusion framework. We perform decision fusion using both maximum and average functions, as illustrated in (7.11) and (7.12), to determine the most appropriate choice between the two. This comprehensive model architecture allows us to capture and integrate complex features from audio and text data effectively.

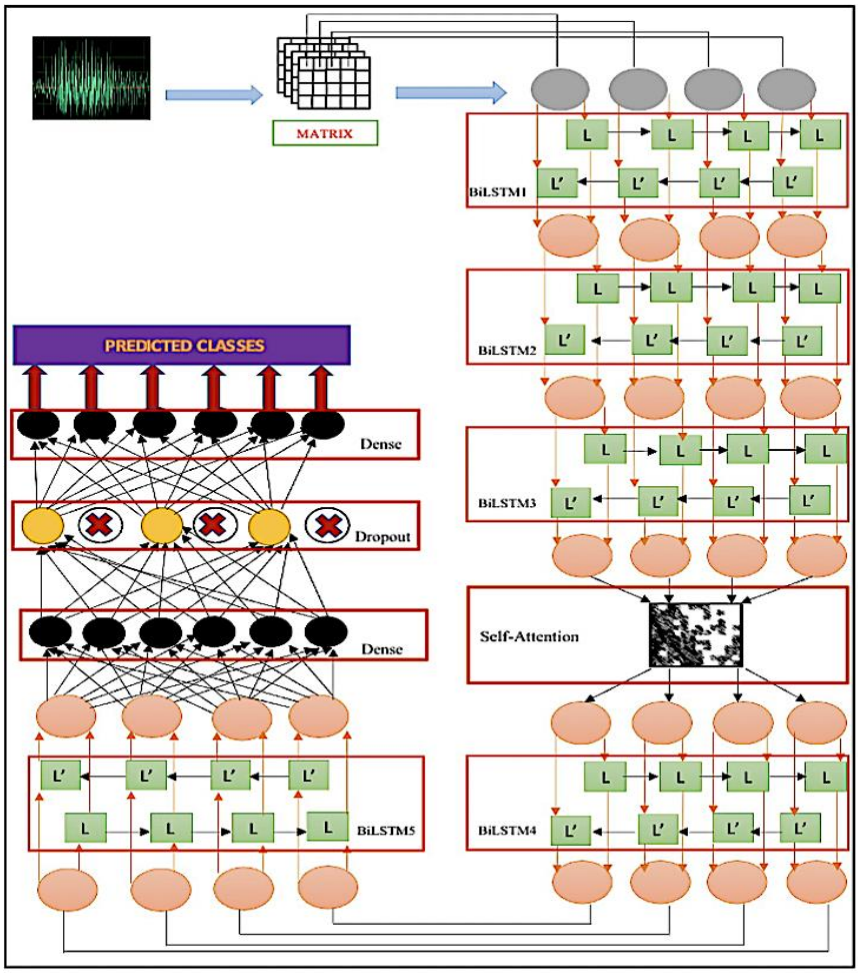


Figure 7.1. Proposed Architecture of Deep BiLSTM with Self-attention

7.2 Experimental Setup And Results

7.2.1 Results And Discussions

The experiments were conducted using Python version 3.9.0 on a macOS High Sierra machine with an Intel Core i5 processor clocked at 1.8GHz. Our code is made accessible online to support future research endeavours. The Librosa library is utilized for extracting MFCC and Mel-spectrogram features. For extracting MFCC features from raw audio files, we use a hop length of 512 and a sampling rate of 22,050 sample points per second. We extract the 13 MFCC features and concatenate them with the delta (1st order) and delta-delta (2nd order) cepstral coefficients. After obtaining 39 MFCC features for each timestamp, standardized to a length of 44, the resulting feature matrix with dimensions 44×39 serves as input to the SA-deep BiLSTM model. In Figure 7.2, you can observe 2D representations of

frame-wise MFCC coefficients computed from raw audio files for the ten word categories of the Google speech command dataset. The Mel-spectrogram is extracted from audio using the Librosa package, following a process involving short-time Fourier transform, conversion of amplitudes to decibels, and further transformation of frequencies to the Mel scale. Figure 7.2 displays 2D representations of the Mel-spectrogram features for samples of the ten word categories from the Google speech command dataset. The Mel-spectrogram feature matrix, with dimensions 44×128, derived from 44 timestamps of each raw audio file, serves as input to the SA-deep BiLSTM model. The trained model generates probabilistic scores for each prediction. We conducted each experiment over 100 epochs, employing the "ReLu" activation function to introduce nonlinearity into the model. We utilized the Adam optimizer for managing the learning rate in stochastic gradient descent. The Adam optimizer is known for its robust performance in various classification tasks, offering fast convergence and a stable learning rate. The loss function employed is sparse categorical cross-entropy. The linguistic characteristics are captured using the text transcript obtained through the Google API. Stemming and conversion of text words into 50-dimensional GloVe word embeddings are performed before feeding them as input to the SA-deep BiLSTM model.

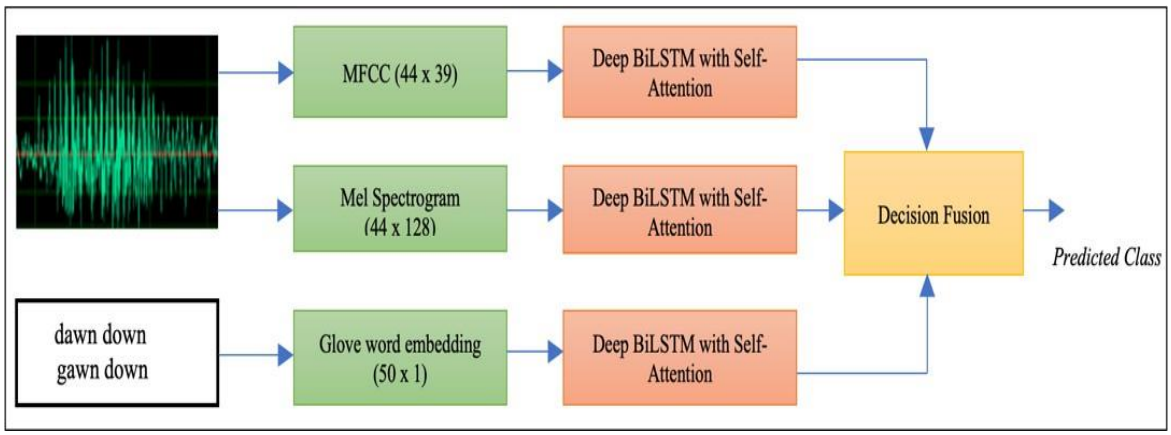


Figure 7.2. Proposed deep fusion framework for speech command recognition using the acoustic and linguistic features

As described in Section 7.1, the posterior class probabilities generated by the three deep models are combined. This comprehensive approach leverages both acoustic and linguistic features for improved speech command recognition. The proposed SA-deep BiLSTM model remains consistent for each of the input modalities within our fusion framework. The

probabilistic scores associated with each class are fused using a soft fusion approach, employing either the average or mean function, as outlined in Eq. (7.11–7.13). The maximum function, as shown by subsequent results, is less effective compared to the average function. The maximum (fused) probability indicates the class of the test sample. In Table 7.1, a presented is presented between the performance of our multimodal fusion approach and SOTA methods. The results reveal that in an adequately resourced data environment, our technique outperforms alternatives, achieving the highest test accuracy of 98.80%. We compared our suggested approach with a CNN that takes a 2D matrix of MFCC features as input. Our technique surpasses the performance of MFCC with CNN by 5.52%. When using the Mel Spectrogram as an input feature, LSTMs are known to perform well in sound classification. However, our suggested method outperforms Mel Spectrogram with LSTM by 3.73%, achieving an accuracy of 95.44%. EdgeCRNN, which utilizes a feature-enhanced method based on residual structure and depth-wise separable convolution, achieves an accuracy of 98.20% on 10 spoken word categories, which is 0.60% less accurate than our method. Our approach also outperforms (Abdelmaksoud et al. 2021), which uses CNN and GFCC as input, by 5.71%. GFCCs are occasionally considered superior signal representations for emotion perception. Additionally, the combination of DenseNet and BiLSTM, as proposed recently by Zeng and Xiao for keyword spotting, demonstrates an accuracy of 94.88%, while our approach surpasses it by 3.92%. These findings underscore the effectiveness of our proposed multimodal fusion approach in speech recognition. In our comparative analysis, we also evaluated our results against the LSTM Architecture explored by Zia and Zahid in for Urdu acoustic modeling. The utterances were pre-processed using the Python Speech Features Toolkit's MFCC approach, with a frame size of 10ms, frame shift of 5ms, 40 filter bank channels, 20 cepstral coefficients, and 58 cepstral parameters. Our approach outperforms this method by 3.66%. For the Google Speech Command dataset, accuracy is presented as the mean of five runs, and the standard deviation, which consistently remains at 0.1%, is not reported. Our method also surpassed the Deep CO-Training algorithm (DCT) by 3.22%. Furthermore, our approach outperformed the attention convolutional RNN by 4.69%. This architecture takes raw .WAV files as inputs, computes mel-scale spectrograms using a non-trainable Keras layer, extracts short- and long-term dependencies, and employs an attention mechanism to determine the most useful information region, which is then fed to a series of dense layers. In a short vocabulary keyword classification challenge, attention-based encoder-decoder models have been shown to outperform baselines,

achieving 97.5% accuracy on TensorFlow's Speech Commands dataset. However, our method surpasses these results by 1.30%. In summary, the results from Table 7.1 demonstrate that our proposed method achieved a high accuracy of 98.80%, outperforming the SOTA for the categorization of the 10-word categories within the Google Speech Command dataset.

Table 7.1 Performance comparison with the state of the art for the 10-word category of Google Speech Command Dataset

Methods	Accuracy (%)
Attention based sequence to sequence model (Higy and Bell 2018)	97.50%
Semi supervised audio tagging **(**Cances and Pellegrini 2021)	95.58%
EdgeCRNN (Wei et al. 2021)	98.20%
RNN Neural attention (de Andrade et al. 2018)	94.11%
DenseNet + BiLSTM (Zeng and Xiao 2018)	94.88%
MFCC + LSTM-RNN (Zia and Zahid 2019)	95.14%
Mel Spectrogram with LSTM (Lezhenin et al. 2019)	95.07%
MFCC + LSTM-RNN (Wazir et al. 2019)	95.44%
GFCC + CNN (Abdelmaksoud et al. 2021)	93.09%
MFCC + CNN (Haque et al. 2020)	93.28%
Proposed Method	98.80%

7.2.3 ABLALTION STUDY

In this analysis, we delve into the impact of individual features, whether acoustic or linguistic, in an ablation study concerning our SA-deep BiLSTM model. The objective is to assess the significance of each component (MFCC, Mel-spectrogram, stem), as well as their pairwise combinations, within our deep fusion framework that employs the SA-deep BiLSTM model. The results of the ablation analysis, summarized in Tables 7.2, and 7.3, have led to the following observations:

Table 7.2 Ablation-analysis of each method with accuracy score obtained for each class

Speech command	MFCC (%)	MS (%)	Stem (%)	OURS (%)
RIGHT	97.47	97.98	64.90	98.23
GO	97.26	95.77	52.99	97.26
NO	99.75	99.01	67.16	99.75
LEFT	99.51	99.76	47.57	100

STOP	99.51	99.76	53.28	99.51
UP	99.76	99.06	27.29	99.53
DOWN	96.06	96.80	49.26	96.80
YES	98.28	99.04	63.96	99.52
ON	98.23	97.73	44.95	98.48
OFF	98.26	98.26	97.26	98.76

Table 7.3 The results of soft fusion by averaging on combination of features for the 10-word Google speech command dataset

Combinatory Results	Type of Combination	Accuracy
Single Component	Stem	56.70%
Single Component	Lemma	56.31%
Single Component	MFCC	98.53%
Single Component	Mel-Spectrogram	98.33%
Two Components	MFCC + LEMMA	98.60%
Two Components	MS + LEMMA	98.43%
Two Components	MFCC + STEM	98.67%
Two Components	MS + STEM	98.53%
Two Components	MFCC + MS	98.70%
Two Components	LEMMA + STEM	62.13%
Three Components	MFCC + LEMMA + STEM	98.64%
Three Components	MS + LEMMA + STEM	98.43%
Three Components	MS + LEMMA + MFCC	98.72%
Three Components	MS + STEM + MFCC	98.80%

Here are the key findings:

- Within our proposed deep framework, audio feature-based classification outperforms text-based classification.
- We have incorporated acoustic and text modalities, specifically MFCC, Mel-spectrogram, and stem, which have demonstrated superior performance compared to combinations of other popular acoustic and linguistic features like GFCC, Log Mel Filterbank, Linear Predictive Cepstral Coefficients (LPCC), and lemma.
- A crucial component in SWR is the self-attention module, integrated between the three initial BiLSTM layers and the two higher-level BiLSTM layers in our SA-deep BiLSTM model. Self-attention serves to emphasize context in the input sequence that is specifically relevant to the classification task at hand.

- The individual accuracy values for MFCC and Mel-Spectrogram were 98.53% and 98.33%, respectively. Through fusion, the accuracy was further enhanced to 98.80%.
- When stemming was applied independently, the word recognition rate was 56.70%. However, incorporating stemming within our fusion framework elevated the accuracy to 98.80%. Substituting stemming with lemmatization resulted in a marginal drop to 98.72%. Lemmatization alone achieved an accuracy of 56.31%, which is slightly lower than the performance of stemming.
- It's evident that our suggested deep-BiLSTM attention strategy following soft fusion achieved nearly 100% accuracy in identifying the speech command "LEFT."
- Several word categories, including "RIGHT," "STOP," "NO," "LEFT," "UP," "YES," "ON," and "OFF," demonstrated high accuracies.
- However, for some word categories, our proposed technique did not yield significant accuracy improvements. Table 3 displays the results of our ablation study, showcasing the accuracy obtained per word category when applying stemming, MFCC, and Mel-spectrogram individually and after decision-level fusion. It's evident that the recognition accuracy for each word category increased significantly due to the soft fusion technique. The high accuracies can also be attributed to the effectiveness of our SA-deep BiLSTM model, which adeptly learns from each modality separately.

Tables 7.2 and 7.3 present the results of various combinations of auditory and linguistic features used in our investigation, both homogeneous and heterogeneous. Morphological text analysis is a common practice in NLP and information retrieval. Stemming and lemmatization are two common morphemes. The stemming algorithm reduces words to their root form or morpheme, thereby supporting vocabulary and text transcript size reduction in information retrieval. In contrast, lemmatization involves removing inflectional endings to reveal the basic structure of a word via morphological analysis. Lemmas carry contextual meaning, whereas stemming focuses on affix removal without considering semantics. The results of soft fusion by averaging (7.13) for various combinations of acoustic/linguistic features are shown in above tables. From our ablation study, we draw the outcomes that improved spoken word categorization can be achieved by combining acoustic and linguistic modalities. Table 7.3 demonstrates that, of the three feature types (MFCC, Mel-Spectrogram, and stem), stemming performed the least effectively due to inaccurate transcriptions from

raw audio files. However, when stem and lemma morphological features are combined, it results in significant improvements of 5.43% and 5.82%, respectively, over their respective baselines. The improvement over separate techniques can be as high as 42.10% for Stem, 42.49% for Lemma, 0.27% for MFCC, and 0.47% for Mel-Spectrogram, for the average metrics score, when MFCC, Mel-Spectrogram, and stem are optimally fused.

7.2.4 IMPLEMENTATION CHALLENGES

Working with multiple modalities, including signals, speech, text, face, and motion data, presents both intriguing possibilities and challenges. One of the primary challenges is the computational complexity involved in training deep networks from scratch on large datasets. Our framework addresses this challenge partially through independent learning for different modalities. It's important to note that the computational complexity is primarily dependent on the length of the input rather than the processing speed of the machine. The operational complexity of our model can be described as follows: the operations for each BiLSTM layer are $O(Lp^2)$, and the self-attention layer is $O(Lp^2)$, where p represents the model dimension of hidden states, and L is the length of the input features. To mitigate this complexity, one potential approach is to use restricted self-attention, although this may come at the cost of reduced accuracy. Additionally, when dealing with larger datasets, it's advisable to work with smaller mini-batch sizes to manage computational demands. Our experiments encountered significant challenges related to error-prone text transcripts obtained from online speech translators like the Google API. These transcripts introduced errors into our data, making the task more challenging. However, the inclusion of the acoustic modality derived from raw audio in our deep framework played a crucial role in mitigating these errors to a considerable extent. Accents represent one of the major hurdles in speech recognition, adding complexity and variability to the spoken language. Furthermore, the diverse nature of speakers, with variations in pronunciation, intonation, and speaking styles, further complicates the task of speech recognition. The presence of a wide range of phonemes, including vowels and diphthongs, in any language can significantly affect pronunciation, translation, word recognition, and keyword tagging. The development of ASR systems can also be impeded by factors such as a lack of diverse training utterances, disorganized speech data, or simple machine errors. These challenges highlight the intricacies of speech recognition and the need for robust approaches to address them.

7.3 SIGNIFICANT OUTCOMES

In this chapter, we present a novel approach involving late fusion of audio and text modalities using the SA-deep BiLSTM model for independent learning of each modality. Our experiments, conducted on the GSCD with 10-word categories, achieved an impressive accuracy of 98.80% by training each modality with a deep self-attention BiLSTM model. We describe a soft fusion method that leverages posterior class probabilities derived from linguistic (stem) and acoustic (MFCC and Mel-spectrogram) features extracted from each audio file. To train these features, we propose the SA-deep BiLSTM model, which comprises five BiLSTM layers and integrates a self-attention module between the third and fourth layers. Our fusion method demonstrates superior classification accuracy compared to the current SOTA techniques for SWR. Notably, our approach excels in correctly predicting the "LEFT" word category. In future investigations, it would be intriguing to explore early-cum-late fusion approaches. Addressing the issue of errors introduced during the Google speech translation process, which leads to the loss of audio-to-text data, could further enhance the results. Additionally, incorporating articulatory features and handling background noise represent potential extensions of this research. Our work harmoniously merges linguistic and acoustic elements, effectively compensating for the shortcomings of each modality. By integrating both acoustic and linguistic information within a deep fusion framework, we achieve more accurate spoken word classification, capturing the valuable insights contributed by both audio and text modalities.

We have conducted additional research aimed at enhancing the accuracy of spoken word recognition by employing a fusion framework that combines raw audio and speech transcriptions.

We explore the potential of combining spectrograms and linguistic data to improve SWR accuracy. Using the GSCD, we extract and resize Mel Spectrogram images for categorization by ImageNet, RegNet, and ConvNext. We employ RegNet and ConvNext architectures, trained on ImageNet, along with the Speech2Text transformer to enhance ASR performance. Probabilistic scores from these models are refined within a deep dense model for spoken utterance classification, augmented by SBERT embeddings for text transcript conversion.

Experiments on the Google Speech Command dataset show impressive accuracy rates, with ConvNext achieving 95.87%, RegNet 99.95%, and text transcripts 85.93%. A late fusion strategy combining word and image embeddings enhances performance.

Our work contributes significantly in the following ways:

- Integration of textual transcripts and mel spectrograms enhances ASR accuracy.
- Transformer-based methods, including RegNet, ConvNext, and Speech2Text, are employed to process audio data and transcripts, leading to remarkable accuracy rates on the Google Speech Command dataset.

The chapter is structured as follows: Section 7.4 outlines our proposed approach and experimental setup, summarizing our methodology and results concisely. Section 7.5 discusses key findings and suggests directions for future research.

7.4 Proposed Approach And Setup

We delve into the unexplored potential of spectrograms and linguistic data to enhance the accuracy of SWR. Our experiments revolve around the GSCD, featuring 35-word categories. Initially, we extract Mel Spectrogram images from audio samples. After resizing these images to 256 x 256 pixels for two-dimensional audio representation, they undergo individual categorization by ImageNet, RegNet, and ConvNext. To boost ASR performance, we leverage RegNet and ConvNext architectures, initially trained on ImageNet's vast dataset of 14 million annotated images. Furthermore, we utilize the Speech2Text transformer to segregate text transcript acquisition from raw audio, generating probabilistic scores. These scores, along with pre-trained ones from RegNet and ConvNext, undergo further refinement within a deep dense model comprising five layers and batch normalization, facilitating spoken utterance classification. Additionally, we employ SBERT embeddings via Siamese BERT-networks to convert Speech2Text transcripts into vectors.

For our experiments, we utilized Google Colab Pro++ with GPU acceleration and Python as the programming language. We employed the Adam optimizer with a learning rate of 0.001 over 100 epochs and a batch size of 32. Nonlinearity was introduced using ReLU activation. Speech-to-image conversion involved generating embeddings and passing them through a densely layered feed-forward network. This network consisted of dense layers with varying neuron counts and batch normalization layers, culminating in 35 output neurons for posterior

probability scores. We used the Adam optimizer and sparse categorical cross-entropy loss function for optimization during training.

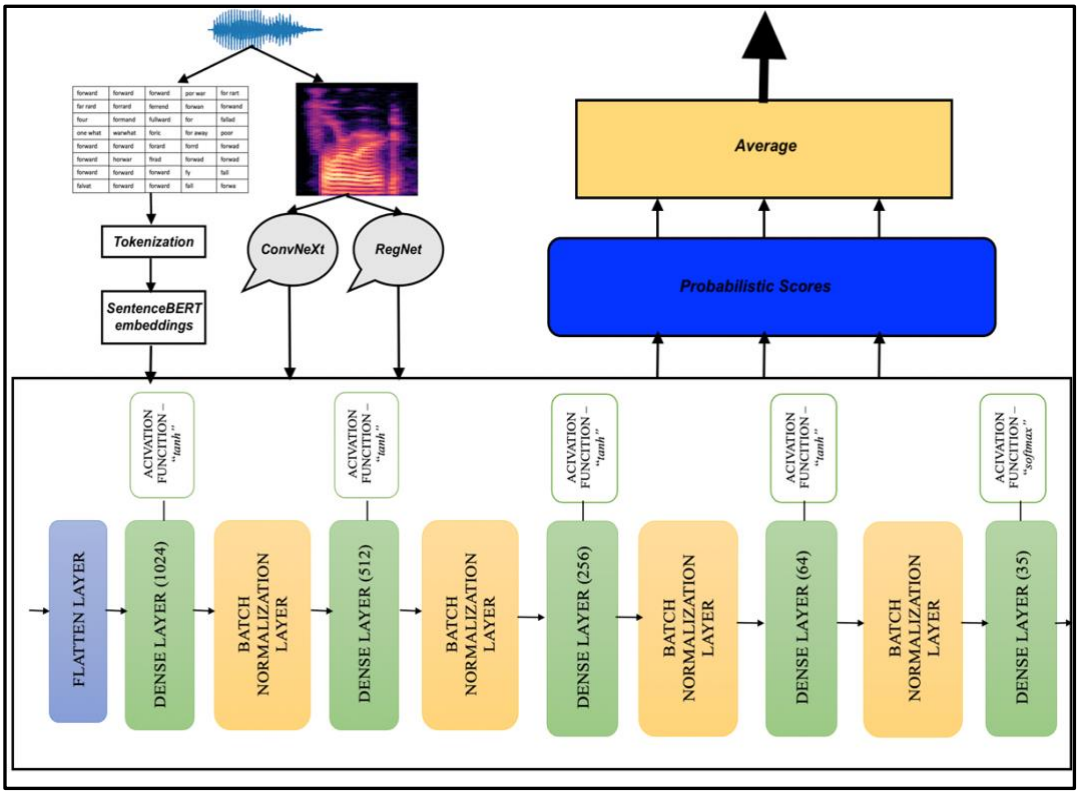


Figure 7.3. The Proposed deep fusion framework

7.5 Significant Outcomes

Our research unveils a novel late fusion strategy that combines word and image embeddings, resulting in remarkable test accuracy rates: 95.87% for ConvNext, 99.95% for RegNet, and 85.93% for text transcripts. Overall, our study not only pushes the boundaries of ASR but also demonstrates the vast potential of integrating multimodal data and advanced transformer-based techniques to achieve exceptional accuracy in spoken word recognition, paving the way for significant advancements in speech recognition systems.

In the next chapter, we have covered the conclusion and outline future directions for our work.

CHAPTER 8

CONCLUSION AND FUTURE WORK

This thesis introduces a fusion-based spoken word classification system designed to assess spoken words and sentences within audio-video datasets. The conceptual framework addresses various challenges in speech analysis, including:

- Insufficient research has been conducted on multiple speech frameworks.
- Limited exploration exists on the concurrent classification of audio-phonemes and text-phonemes.
- There is a gap in the existing literature regarding the joint extraction and classification of features from both raw audio and speech transcription.
- The simultaneous classification of phonemes and morphemes has been an underexplored area in current research.

A novel supervised fusion-based system has been developed for multimodal speech classification, proficient in identifying speech through the integration of both acoustic and linguistic information. This system incorporates five innovative fusion rules, leveraging the joint utilization of linguistic and acoustic features. In conclusion, this thesis addresses the pressing need for advanced fusion-based algorithms in Speech Analysis. The challenges identified in spoken word analysis, encompassing phonological and morphological aspects, prompt the development of innovative solutions across seven chapters. The proposed fusion-based frameworks offer nuanced approaches to enhance spoken word classification and recognition, demonstrating their effectiveness in handling the inherent ambiguities of natural language. Each chapter presents a unique solution to a specific problem:

Phonological Studies and Spectrogram-based Spoken Word Classification (Chapter 3): Novel dual-pronged approach integrating spectrograms and phonology. Speech2Text transformer for text transcript extraction and spectrogram generation. Late fusion

strategy combining phone and image embeddings. Achieved benchmark accuracy, surpassing existing methods. Phoneme-Based SWR (Chapter 4): Supervised approach for accented speech recognition using phonology. Early fusion of phone embeddings to identify accented speech. Superior accuracy in the context of the L2-ARCTIC accented speech corpus. Enhanced Speech Transcription Accuracy (Chapter 5): Unsupervised approach for improving imperfect speech transcriptions. Decision-level fusion involving stemming and two-way phoneme pruning. Significant improvement in word recognition rates on the LRW dataset. Morphological and Phonological Classification (Chapter 6): Supervised technique for SWR using phonemes and bigrams. Late fusion approach incorporating phone and bigram embeddings. Improved accuracy in Arabic, Vietnamese, and Tamil for multilingual spoken words. Multimodal Data Integration (Chapter 7): Late fusion of audio and text modalities for deep speech recognition. SA-deep BiLSTM model processes audio and text data independently. Impressive accuracy on the Google speech command dataset, outperforming SOTA approaches.

In the pursuit of these solutions, each chapter contributes to advancing the field of Speech Analysis by embracing fusion-based strategies. The proposed frameworks exhibit versatility in handling diverse challenges, from phonological studies to ML SWR. Overall, this thesis establishes the efficacy of fusion-based algorithms in comprehensively addressing the complexities of spoken word analysis.

In the concluding remarks of Chapter 3, we introduced three innovative approaches inspired by spectrogram and phonological stress-level studies. Our study highlights the significant role played by vowels and plosives in speech recognition, with stress markers notably contributing to transcription accuracy. By focusing on these crucial phonological features, further advancements in speech recognition technologies can be achieved, catering to a broad range of users and languages. Among the various sounds analysed, trills, characterized by rapid articulator vibrations, demonstrated the second-highest accuracy. Integration of stress markers through Sentence BERT resulted in a test accuracy of 70.81% for vowels and trills, while the USE achieved 72.03%. In the absence of stress markers, accuracy dropped to 62.13% (Sentence BERT) and 70.64% USE. Interestingly, stress markers proved instrumental in enhancing accuracy, with the USE outperforming Sentence BERT in this specific context. To optimize results, future research should delve into trill variations across different languages and refine the incorporation of stress markers for improved speech recognition in diverse linguistic

contexts. Nasals, which involve consonant sounds produced by allowing air to pass through the nose while obstructing the oral cavity, exhibited noteworthy accuracy in the assessed sounds. Integration of stress markers through Sentence BERT resulted in a test accuracy of 66.29% for vowels and nasals, while the USE achieved 53.01%. Without stress markers, accuracy remained relatively high at 65.92% (Sentence BERT) and 66.55% USE. Surprisingly, despite the general trend of stress markers improving accuracy, the USE outperformed Sentence BERT in this instance. Remarkably, the USE's embeddings, generated without considering stress markers, proved to be the most accurate. To enhance performance, future investigations should delve into the distinctive aspects of nasals in various languages and optimize the utilization of stress markers to refine speech recognition across diverse linguistic contexts.

This thesis presents a novel approach to boost speech command recognition by employing a late fusion technique that combines audio and image modalities. The methodology incorporates a feed-forward neural network model along with median and adaptive filtering techniques to enhance audio signals. Employing the GSCD featuring 10-word categories, the approach achieves an impressive test accuracy of 99.85% by integrating information from both audio and image modalities. The fusion technique employs soft fusion, utilizing posterior class probabilities obtained from two filtered channels derived from each audio file. The feature extraction process involves the Swin-Tiny Transformer, followed by a 3-layered feed-forward neural network. In comparison to existing SOTA methods, the proposed fusion technique excels in classification accuracy, showcasing its efficacy in capturing and utilizing information from both audio and image modalities for detailed speech command classification. The integration of a pre-trained Swin-Tiny Transformer model, trained on an extensive image dataset, significantly contributes to the achieved high accuracy. Notably, median filtering emerges as a superior pre-processing technique when compared to adaptive filtering. The fusion of probabilistic outcomes from adaptive and median filtering further enhances accuracy, resulting in an exceptional success rate of 99.85%. It's important to note that one limitation is the increased memory demand, particularly when applied to extensive datasets, a challenge shared across similar research endeavours using the same dataset.

This research presents an extensive statistical analysis of deep-feature-level fusion involving audio and visual modalities through an innovative deep BiLSTM-GRU model,

dedicated to learning each modality. Leveraging training output scores from audio and image modalities, our approach attains notable accuracies: 98.56% for 20 utterances of dysarthric male speakers, 98.07% for 10 utterances of dysarthric male speakers, 95.11% for 20 utterances of dysarthric female speakers, 94.87% for 10 utterances of dysarthric female speakers, 97.55% for 20 utterances of both male and female dysarthric speakers, and 94.80% for 10 utterances of both male and female dysarthric speakers, utilizing 10- and 20-word categories from the Dysarthria speech corpus. Our proposed feature-level fusion method incorporates transformers-based feature engineering, ultimately generating posterior class probabilities for both audio and image modalities. Despite the promising results, there remain various avenues for further exploration. Future work will involve integrating linguistic aspects into the multi-modal feature fusion module, ongoing refinement of feature extraction methodologies, and extending our model to encompass aspect-based multi-modal speech recognition. Additionally, exploration of other multimedia data types, such as text, will be pursued.

In Chapter 4, this research investigates the categorization of low-resourced accented speech by employing an innovative approach involving the early fusion of phoneme embeddings and a dense model. Notably, our method breaks new ground by combining early phone fusion with accented speech, representing a novel initiative within the realm of low-resourced datasets. The integration of early fusion of phoneme embeddings with a 3-layered dense model significantly enhances the performance of recognizing low-resourced accented speech. Our experiments were conducted using the L2-ARCTIC accented speech dataset, which includes audio recordings from 24 non-native English speakers across 10 sentence categories. The proposed methodology has yielded commendable results in the classification of spoken sentences. Our future research will concentrate on further exploring diverse fusion models tailored to low-resourced accented speech corpora, thereby expanding the scope of this study.

In Chapter 5, our methodology revolves around combining two potent techniques, stemming and two-way phoneme pruning, to enhance word recognition accuracy in highly imperfect speech transcriptions extracted from the LRW dataset in mp4 format. The process begins with the extraction of audio samples from videos using the Ffmpeg framework, followed by converting the audio speech into text transcriptions using the publicly available Google API, versatile in applications such as speech adaption, speech transcription, and real-time speech recognition. To assess our results, we start with a

baseline comparison involving simple string matching to identify word categories in the text transcription. Our initial step involves text normalization and speech adaption, which includes removing stop words—the most frequent and extraneous words in the text—to expedite text processing. Subsequently, we apply stemming to derive the root form of each word, comparing it against various word categories. Simultaneously, we convert each word into phonemes using the CMU pronouncing dictionary. The text transcript is then mapped to phonemes, followed by phoneme filtering where we selectively filter out phonemes containing vowels, plosives, or fricatives. The phoneme pruning process consists of two non-sequential stages: Stage I involves phoneme pruning using vowels and plosives, while Stage II focuses on phoneme pruning using vowels and fricatives. We then aggregate results from these three methods and apply decision fusion to determine whether any of these methods successfully detect the occurrence of the word. The proposed fusion method proves highly effective, surpassing existing SOTA techniques. Consequently, word recognition accuracy sees a significant improvement, elevating it from a baseline accuracy of 9.34% to an impressive 32.96% using our fusion approach.

In Chapter 6, this investigation delves into the classification of low-resource single-word audio datasets through an innovative approach: a dense model created by the late fusion of phoneme embeddings and bigram embeddings. What distinguishes our method is the application of late fusion to a single voice dataset with limited resources, a novel approach in itself. We observed a significant enhancement in the performance of the low-resource keyword spotting dataset when integrating phoneme embeddings and bigram embeddings into a 5-layered dense model with batch normalization. Our experiments employed the MSWC dataset, comprising natural speaker audio recordings and ten-word categories. The outcomes from our proposed approach for spoken word classification were promising. Interestingly, we noted that text transcripts can exert a substantial impact on spoken word classification, surpassing audio-based features. We conducted comparisons with existing SOTA methods and carried out an ablation analysis of various NLP techniques for the selected Asian languages.

In another study, we presented a supervised technique for SWR in a ML corpus with limited resources. A significant gap in the current state of the art lies in the insufficient exploration of morphology and phonology in deciphering spoken text. For the 10 spoken

word categories in the MSWC, the early fusion of phone embeddings and unigrams embeddings yielded accuracies of 68.96% and 44.83%, respectively. In the case of 50 spoken word categories, these accuracies were 67.28% and 52.78%. Following early fusion, the late fusion of probabilistic scores from both approaches was applied. In comparison to prior methods, the late fusion of phone embeddings and unigrams embeddings for SWR achieved accuracies of 72.41% and 70.97% for 10 and 50 spoken word categories in the MSWC, respectively. Our research aims to underscore that even with a limited number of training samples, phonemes derived from raw audio can significantly impact MSWC. Future work includes expanding our research to more native languages and exploring the potential role of audio phonemes in spoken word recognition.

In Chapter 7, we introduce a ground-breaking method that involves the late fusion of audio and text modalities using the SA-deep BiLSTM model for independent learning of each modality. Our experiments, conducted on the GSCD with 10-word categories, yielded an impressive accuracy of 98.80% by training each modality with a deep self-attention BiLSTM model. We detail a soft fusion technique that utilizes posterior class probabilities derived from linguistic (stem) and acoustic (MFCC and Mel-spectrogram) features extracted from each audio file. To train these features, we propose the SA-deep BiLSTM model, consisting of five BiLSTM layers and integrating a self-attention module between the third and fourth layers. Our fusion method showcases superior classification accuracy compared to current SOTA techniques for SWR. Notably, our approach excels in correctly predicting the "LEFT" word category. In future investigations, exploring early-cum-late fusion approaches could be intriguing. Addressing errors introduced during the Google speech translation process, leading to the loss of audio-to-text data, could further enhance the results. Additionally, incorporating articulatory features and addressing background noise represent potential extensions of this research. Our work seamlessly combines linguistic and acoustic elements, effectively compensating for the shortcomings of each modality. By integrating both acoustic and linguistic information within a deep fusion framework, we achieve more accurate spoken word classification, capturing valuable insights contributed by both audio and text modalities.

LIST OF PUBLICATIONS

Journal Publications

1. Sunakshi Mehra, Seba Susan. "Deep fusion framework for speech command recognition using acoustic and linguistic features. Multimedia Tools and Applications (2023): Vol 82." 38667–38691 (2023). (**SCIE-Indexed, IF: 3.6**). https://doi.org/10.1007/s11042-023-15118-1
2. Sunakshi Mehra, Virender Ranga, & Ritu Agarwal. "Improving speech command recognition through decision-level fusion of deep filtered speech cues." Signal, Image and Video Processing SIViP (2023). (**SCIE-Indexed, IF: 2.3**). https://doi.org/10.1007/s11760-023-02845-z
3. Sunakshi Mehra, Virender Ranga and Ritu Agarwal. "A deep learning approach to dysarthric utterance classification with BiLSTM-GRU, speech cue filtering, and log mel spectrograms." accepted in The Journal of Supercomputing (2023)". (**SCIE-Indexed, IF: 3.3**). https://doi.org/10.1007/s11227-024-06015-x
4. Sunakshi Mehra, Virender Ranga, Ritu Agarwal and Seba Susan. "Speaker independent recognition of low-resourced multilingual Arabic spoken words through hybrid fusion." accepted in Multimedia Tools and Applications (2023). (**SCIE-Indexed, IF: 3.6**). https://doi.org/10.1007/s11042-024-18804-w
5. Sunakshi Mehra, Virender Ranga and Ritu Agarwal. "Multimodal integration of mel spectrograms and text transcripts for enhanced automatic speech recognition: Leveraging extractive transformer-based approaches and late fusion strategies" currently under review in Computational Intelligence (2024). (**SCIE-Indexed, IF: 2.1**).
6. Sunakshi Mehra, Virender Ranga and Ritu Agarwal. "Evaluating the significance of suprasegmental features in speech command recognition through spectrogram and phonological fusion analysis" currently with editor in Soft Computing (2024)". (**SCIE-Indexed, IF: 3.1**).
7. Sunakshi Mehra, Virender Ranga, Ritu Agarwal and Seba Susan. "Spoken Word Recognition for Asian Languages using Transformers" currently under review in Computer Speech and Language (2023). (**SCIE-Indexed, IF: 4.3**).
8. Sunakshi Mehra, Virender Ranga and Ritu Agarwal. "Dhivehi Speech Recognition: A Multimodal Approach for Dhivehi Language in Resource-Constrained Settings" under review in Circuits, Systems, and Signal Processing. (**SCIE-Indexed, IF: 1.8**).
9. Sunakshi Mehra, Virender Ranga, Ritu Agarwal, & Seba Susan. "Phonetic and Lexical Modeling for Speaker-Independent Automatic Speech Recognition: A Contemporary Study" under preparation.

Conferences

10. Sunakshi Mehra, and Seba Susan. "Early Fusion of Phone Embeddings for Recognition of Low-Resourced Accented Speech." In 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), pp. 1-5. IEEE, 2022. https://doi.org/10.1109/AIST55798.2022.10064735
11. Sunakshi Mehra, & Seba Susan (2021). Improving word recognition in speech transcriptions by decision-level fusion of stemming and two-way phoneme pruning. In Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10 (pp. 256-266). Springer Singapore. https://doi.org/10.1007/978-981-16-0401-0_19

BIBLIOGRAPHY

Abdelmaksoud, E. R., Hassen, A., Hassan, N., & Hesham, M. (2021). Convolutional neural network for arabic speech recognition. *The Egyptian Journal of Language Engineering*, *8*(1), 27-38. https://doi.org/10.21608/ejle.2020.47685.1015

Abdou, S. M., & Moussa, A. M. (2019). Arabic speech recognition: Challenges and state of the art. *Computational linguistics, speech and image processing for arabic language*, 1-27. https://doi.org/10.1142/9789813229396_0001

AbuZeina, D., Al-Khatib, W., Elshafei, M., & Al-Muhtaseb, H. (2011). Cross-word Arabic pronunciation variation modeling for speech recognition. *International Journal of Speech Technology*, *14*, 227-236. https://doi.org/10.1007/s10772-011-9098-0

Alghamdi, M. (2001). Arabic phonetics. *Al-Toubah Bookshop, Riyadh*.

Alsharhan, E., & Ramsay, A. (2019). Improved Arabic speech recognition system through the automatic generation of fine-grained phonetic transcriptions. *Information Processing & Management*, *56*(2), 343-353. https://doi.org/10.1016/j.ipm.2017.07.002

Alsharhan, E., & Ramsay, A. (2020). Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition. *Language Resources and Evaluation*, *54*, 975-998. https://doi.org/10.1007/s10579-020-09505-5

Alsulaiman, M., Mahmood, A., & Muhammad, G. (2017). Speaker recognition based on Arabic phonemes. *Speech Communication*, *86*, 42-51. https://doi.org/10.1016/j.specom.2016.11.004

Alzahra, B., Ko, K., & Ko, H. (2019). Bird sounds classification by combining PNCC and robust Mel-log filter bank features. *The Journal of the Acoustical Society of Korea*, *38*(1), 39-46. https://doi.org/10.7776/ASK.2019.38.1.039

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Zhu, Z. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning* (pp. 173-182). PMLR. https://doi.org/10.48550/arXiv.1512.02595

Arısoy, E., Sak, H., & Saraçlar, M. (2007). Language modeling for automatic Turkish broadcast news transcription. In *Eighth Annual Conference of the International Speech Communication Association*. https://doi.org/10.21437/Interspeech.2007-273

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*. https://doi.org/10.48550/arXiv.1607.06450

Baber, C. (1991). Speech technology in control room systems: A human factors perspective. (1991).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. https://doi.org/10.48550/arXiv.1409.0473

Bastanfard, A., Amirkhani, D., & Naderi, S. (2020, December). A singing voice separation method from Persian music based on pitch detection methods. In *2020 6th Iranian conference*

on signal processing and intelligent systems (ICSPIS) (pp. 1-7). IEEE. https://doi.org/10.1109/ICSPIS51611.2020.9349583

Beguš, G. (2020). Modeling unsupervised phonetic and phonological learning in Generative Adversarial Phonology. *Proceedings of the Society for Computation in Linguistics*, *3*(1), 138-148. https://doi.org/10.48550/arXiv.2006.03965

Berg, A., O'Connor, M., & Cruz, M. T. (2021). Keyword transformer: A self-attention model for keyword spotting. *arXiv preprint arXiv:2104.00769*. https://doi.org/10.48550/arXiv.2104.00769

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, *56*, 85-100. https://doi.org/10.1016/j.specom.2013.07.008

Bhatt, S., Dev, A., & Jain, A. (2020). Confusion analysis in phoneme based speech recognition in Hindi. *Journal of Ambient Intelligence and Humanized Computing*, *11*, 4213-4238. https://doi.org/10.1007/s12652-020-01703-x

Billa, J. (2018, September). ISI ASR System for the Low Resource Speech Recognition Challenge for Indian Languages. In *INTERSPEECH* (pp. 3207-3211). https://doi.org/10.21437/Interspeech.2018-2473

Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, *50*(5), 434-451. https://doi.org/10.1016/j.specom.2008.01.002

Bisol, L. (Ed.). (2005). *Introdução a estudos de fonologia do português brasileiro*. EdiPUCRS. https://doi.org/10.1590/S0102-44502000000100008

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135-146. https://doi.org/10.48550/arXiv.1607.04606

Borges, M. V. (2000). Fonética e fonologia do português: roteiro de estudos e guia de exercícios. https://doi.org/10.1590/S0102-44502000000100011

Boumehdi, A., & Yousfi, A. (2022). Arabic speech recognition independent of vocabulary for isolated words. In *Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 3* (pp. 585-595). Springer Singapore. https://doi.org/10.1007/978-981-16-1781-2_52

Byrd, R. H., Chin, G. M., Neveitt, W., & Nocedal, J. (2011). On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, *21*(3), 977-995. https://doi.org/10.1137/10079923X

Byrne, W., Hajič, J., Ircing, P., Krbec, P., & Psutka, J. (2000, September). Morpheme based language models for speech recognition of Czech. In *International Workshop on Text, Speech and Dialogue* (pp. 211-216). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45323-7_36

Cabrera, R., Liu, X., Ghodsi, M., Matteson, Z., Weinstein, E., & Kannan, A. (2021). Language model fusion for streaming end to end speech recognition. *arXiv preprint arXiv:2104.04487*. https://doi.org/10.48550/arXiv.2104.04487

Cances, L., & Pellegrini, T. (2021, June). Comparison of Deep Co-Training and Mean-Teacher approaches for semi-supervised audio tagging. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 361-365). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9415116

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Kurzweil, R. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*. https://doi.org/10.48550/arXiv.1803.11175

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., & Dubnov, S. (2022, May). HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 646-650). IEEE. https://doi.org/10.48550/arXiv.2202.00874

Chen, Z. (2022, November). Noise Reduction of Bird Calls Based on a Combination of Spectral Subtraction, Wiener Filtering, and Kalman Filtering. In *2022 IEEE 5th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)* (pp. 512-517). IEEE. https://doi.org/10.1109/AUTEEE56487.2022.9994282

Creutz, M., & Lagus, K. (2005). *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki: Helsinki University of Technology. https://doi.org/10.1007/11780885_34

Damper, R. I. (1982). Speech technology—implications for biomedical engineering. *Journal of Medical Engineering & Technology*, *6*(4), 135-149. https://doi.org/10.3109/03091908209041006

Daniel, J., & James, H. (2009). Martin, Speech and Language Processing.

Das, P., & Bhattacharjee, U. (2014, July). Robust speaker verification using GFCC and joint factor analysis. In *Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICCCNT.2014.6963092

Das, R., & Singh, T. D. (2023). Image–Text Multimodal Sentiment Analysis Framework of Assamese News Articles Using Late Fusion. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(6), 1-30. https://doi.org/10.1145/3584861

Dash, T. K., & Solanki, S. S. (2020). Speech intelligibility based enhancement system using modified deep neural network and adaptive multi-band spectral subtraction. *Wireless Personal Communications*, *111*, 1073-1087. https://doi.org/10.1007/s11277-019-06902-0

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357-366. https://doi.org/10.1109/TASSP.1980.1163420

De Andrade, D. C., Leo, S., Viana, M. L. D. S., & Bernkopf, C. (2018). A neural attention model for speech command recognition. *arXiv preprint arXiv:1808.08929*. https://doi.org/10.48550/arXiv.1808.08929

Dede, G., & Sazlı, M. H. (2010). Speech recognition with artificial neural networks. *Digital Signal Processing*, *20*(3), 763-768. https://doi.org/10.1016/j.dsp.2009.10.004

Deller Jr, J. R. (1993). Discrete-time processing of speech signals. In *Discrete-time processing of speech signals* (pp. 908-908). https://doi.org/10.1109/9780470544402

Dhanalakshmi, M., Mariya Celin, T. A., Nagarajan, T., & Vijayalakshmi, P. (2018). Speech-input speech-output communication for dysarthric speakers using HMM-based speech recognition and adaptive synthesis system. *Circuits, Systems, and Signal Processing*, *37*, 674-703. https://doi.org/10.1007/s00034-017-0567-9

Dokuz, Y., & Tüfekci, Z. (2022). Feature-based hybrid strategies for gradient descent optimization in end-to-end speech recognition. *Multimedia Tools and Applications*, *81*(7), 9969-9988. https://doi.org/10.1007/s11042-022-12304-5

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, *12*(7). https://doi.org/10.5555/1953048.2021068

Enderby, P. (2013). Disorders of communication: dysarthria. *Handbook of clinical neurology*, *110*, 273-281. https://doi.org/10.1016/B978-0-444-52901-5.00022-8

Erdogan, H., Buyuk, O., & Oflazer, K. (2005, November). Incorporating language constraints in sub-word based speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* (pp. 98-103). IEEE. https://doi.org/10.1109/ASRU.2005.1566516

Esfe, M. H., Kamyab, M. H., & Toghraie, D. (2022). Statistical review of studies on the estimation of thermophysical properties of nanofluids using artificial neural network (ANN). *Powder Technology*, *400*, 117210. https://doi.org/10.1016/j.powtec.2022.117210

Fang, A., Filice, S., Limsopatham, N., & Rokhlenko, O. (2020, July). Using phoneme representations to build predictive models robust to ASR errors. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 699-708). https://doi.org/10.1145/3397271.3401050

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*. https://doi.org/10.48550/arXiv.2007.01852

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*. https://doi.org/10.48550/arXiv.2103.15122

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, *12*(2), 75-98. https://doi.org/10.1006/csla.1998.0043

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, *17*(1), 2096-2030. https://doi.org/10.48550/arXiv.1505.07818

Gao, Y., Srivastava, B. M. L., & Salsman, J. (2018, May). Spoken english intelligibility remediation with pocketsphinx alignment and feature extraction improves substantially over the state of the art. In *2018 2nd IEEE Advanced Information Management, Communicates,*

Electronic and Automation Control Conference (IMCEC) (pp. 924-927). IEEE. https://doi.org/10.1109/IMCEC.2018.8469649

Gazneli, A., Zimerman, G., Ridnik, T., Sharir, G., & Noy, A. (2022). End-to-end audio strikes back: Boosting augmentations towards an efficient audio classification network. *arXiv preprint arXiv:2204.11479*. https://doi.org/10.48550/arXiv.2204.11479

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017, July). Convolutional sequence to sequence learning. In *International conference on machine learning* (pp. 1243-1252). PMLR. https://doi.org/10.48550/arXiv.1705.03122

Gerkmann, T., & Vincent, E. (2018). Spectral masking and filtering. *Audio source separation and speech enhancement*, 65-85. https://doi.org/10.1002/9781119279860.ch5

Ghoshal, A., Swietojanski, P., & Renals, S. (2013, May). Multilingual training of deep neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 7319-7323). IEEE. https://doi.org/10.1109/ICASSP.2013.6639084

Glocker, K., Herygers, A., & Georges, M. (2023). Allophant: Cross-lingual Phoneme Recognition with Articulatory Attributes. *arXiv preprint arXiv:2306.04306*. https://doi.org/10.48550/arXiv.2306.04306

Gong, Y., Chung, Y. A., & Glass, J. (2021). Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*. https://doi.org/10.48550/arXiv.2104.01778

Gonzalez, R. C., & Woods, R. E. (2008). Digital image processing, prentice hall. *Upper Saddle River, NJ*. https://doi.org/10.1007/3-540-27563-0

Goronzy, S., Rapp, S., & Kompe, R. (2004). Generating non-native pronunciation variants for lexicon adaptation. *Speech Communication*, *42*(1), 109-123. https://doi.org/10.1016/j.specom.2003.09.003

Gupta, P. (2020, February). A context-sensitive real-time Spell Checker with language adaptability. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)* (pp. 116-122). IEEE. https://doi.org/10.48550/arXiv.1910.11242

Hacioglu, K., Pellom, B. L., Ciloglu, T., Öztürk, Ö., Kurimo, M., & Creutz, M. (2003, September). On lexicon creation for turkish LVCSR. In *INTERSPEECH*. https://doi.org/10.21437/Eurospeech.2003-378

Haridas, A. V., Marimuthu, R., & Chakraborty, B. (2018). A novel approach to improve the speech intelligibility using fractional delta-amplitude modulation spectrogram. *Cybernetics and Systems*, *49*(7-8), 421-451. https://doi.org/10.1080/01969722.2018.1448241

Haque, M. A., Verma, A., Alex, J. S. R., & Venkatesan, N. (2020). Experimental evaluation of CNN architecture for speech recognition. In *First International Conference on Sustainable Technologies for Computational Intelligence: Proceedings of ICTSCI 2019* (pp. 507-514). Springer Singapore. https://doi.org/10.1007/978-981-15-0029-9_40

Hawley, M. S., Cunningham, S. P., Green, P. D., Enderby, P., Palmer, R., Sehgal, S., & O'Neill, P. (2012). A voice-input voice-output communication aid for people with severe speech impairment. *IEEE Transactions on neural systems and rehabilitation engineering*, *21*(1), 23-31. https://doi.org/10.1109/TNSRE.2012.2209678

Hawley, M. S., Enderby, P., Green, P., Cunningham, S., Brownsell, S., Carmichael, J., ... & Palmer, R. (2007). A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics*, *29*(5), 586-593. https://doi.org/10.1016/j.medengphy.2006.06.009

Haubold, A., & Kender, J. R. (2007, June). Introduction Of Video Journals And Archives In The Classroom. In *2007 Annual Conference & Exposition* (pp. 12-985). https://doi.org/10.21437/Eurospeech.2003-378

Hazen, T. J. (2006). Automatic alignment and error correction of human generated transcripts for long speech recordings. In *Ninth International Conference on Spoken Language Processing*. https://doi.org/10.21437/Interspeech.2006-449

He, N., & Ferguson, S. (2022). Music emotion recognition based on segment-level two-stage learning. *International Journal of Multimedia Information Retrieval*, *11*(3), 383-394. https://doi.org/10.1007/s13735-022-00230-z

He, X., & Deng, L. (2022). *Discriminative learning for speech recognition: theory and practice*. Springer Nature. https://doi.org/10.1007/978-3-031-02557-0

Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M. A., Devin, M., & Dean, J. (2013, May). Multilingual acoustic models using distributed deep neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8619-8623). IEEE. https://doi.org/10.1109/ICASSP.2013.6639348

Herbig, T., Gerl, F., Minker, W., & Haeb-Umbach, R. (2011). Adaptive systems for unsupervised speaker tracking and speech recognition. *Evolving Systems*, *2*, 199-214. https://doi.org/10.1007/s12530-011-9034-1

Hermansky, H. (1997, December). The modulation spectrum in the automatic recognition of speech. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings* (pp. 140-147). IEEE. https://doi.org/10.1109/ASRU.1997.658998

Higy, B., & Bell, P. (2018). Few-shot learning with attention-based sequence-to-sequence models. *arXiv preprint arXiv:1811.03519*. https://doi.org/10.48550/arXiv.1811.03519

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, *29*(6), 82-97. https://doi.org/10.1109/MSP.2012.2205597

Hinsvark, A., Delworth, N., Del Rio, M., McNamara, Q., Dong, J., Westerman, R., ... & Jette, M. (2021). Accented speech recognition: A survey. *arXiv preprint arXiv:2104.10747*. https://doi.org/10.48550/arXiv.2104.10747

Hirsimäki, Teemu, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen. "Unlimited vocabulary speech recognition with morph language models applied to Finnish." *Computer Speech & Language* 20, no. 4 (2006): 515-541. https://doi.org/10.1016/j.csl.2005.07.002

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Horga, D. (1999). Handbook of the international phonetic association. a guide to the use of the international phonetic alphabet cambridge: Cambridge university press (1999),(204 stranice). *Govor*, *16*(2), 181-188. https://doi.org/10.1017/S0952675700003894

Hu, X., Saiko, M., & Hori, C. (2014, December). Incorporating tone features to convolutional neural network to improve Mandarin/Thai speech recognition. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific* (pp. 1-5). IEEE. https://doi.org/10.1109/APSIPA.2014.7041576

Huggins-Daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., & Rudnicky, A. I. (2006, May). Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *2006 IEEE international conference on acoustics speech and signal processing proceedings* (Vol. 1, pp. I-I). IEEE. https://doi.org/10.1109/ICASSP.2006.1659988

Humphries, J. J., Woodland, P. C., & Pearce, D. (1996, October). Using accent-specific pronunciation modelling for robust speech recognition. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 4, pp. 2324-2327). IEEE. https://doi.org/10.1109/ICSLP.1996.607273

Iqbal, A., & Aftab, S. (2020). A Classification Framework for Software Defect Prediction Using Multi-filter Feature Selection Technique and MLP. *International Journal of Modern Education & Computer Science*, *12*(1). https://doi.org/10.5815/ijmecs.2020.01.03

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv 2015. *arXiv preprint arXiv:1502.03167*. https://doi.org/10.48550/arXiv.1502.03167

Jayaram, G., & Abdelhamied, K. (1995). Experiments in dysarthric speech recognition using artificial neural networks. *Journal of rehabilitation research and development*, *32*, 162-162.

Jacobs, I. S., & Bean, C. P. (1963). Fine particles, thin films and exchange anisotropy (effects of finite dimensions and interfaces on the basic properties of ferromagnets). *Spin arrangements and crystal structure, domains, and micromagnetics*, *3*, 271-350. https://doi.org/10.1016/B978-0-12-575303-6.50013-0

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, *3*(1), 79-87. https://doi.org/10.1162/neco.1991.3.1.79

Jain, A., Singh, V. P., & Rath, S. P. (2019). A Multi-Accent Acoustic Model Using Mixture of Experts for Speech Recognition. In *Interspeech* (pp. 779-783). https://doi.org/10.21437/Interspeech.2019-1667

Jain, A., Upreti, M., & Jyothi, P. (2018, September). Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning. In *Interspeech* (pp. 2454-2458). https://doi.org/10.21437/Interspeech.2018-1864

Janet, H., & Nick, W. (2017). An Introduction to Sociolinguistics (Learning about Language). https://doi.org/10.4324/9780367821852

Kanisha, B., Mahalakshmi, V., Baskar, M., Vijaya, K., & Kalyanasundaram, P. (2022). Smart communication using tri-spectral sign recognition for hearing-impaired people. *The Journal of Supercomputing*, 1-14. https://doi.org/10.1007/s11227-021-03968-1

Kannan, A., Datta, A., Sainath, T. N., Weinstein, E., Ramabhadran, B., Wu, Y., ... & Lee, S. (2019). Large-scale multilingual speech recognition with a streaming end-to-end model. *arXiv preprint arXiv:1909.05330*. https://doi.org/10.48550/arXiv.1909.05330

Khaleelur Rahiman, P. F., Jayanthi, V. S., & Jayanthi, A. N. (2021). Retracted: Speech enhancement method using deep learning approach for hearing-impaired listeners. *Health informatics journal*, *27*(1), 1460458219893850. https://doi.org/10.1177/1460458220943995

Kherdekar, V. A., & Naik, S. A. (2021). Convolution neural network model for recognition of speech for words used in mathematical expression. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(6), 4034-4042. https://doi.org/10.17762/turcomat.v12i6.8374

Kibria, S., Samin, A. M., Kobir, M. H., Rahman, M. S., Selim, M. R., & Iqbal, M. Z. (2022). Bangladeshi Bangla speech corpus for automatic speech recognition research. *Speech Communication*, *136*, 84-97. https://doi.org/10.1016/j.specom.2021.12.004

Kim, C., & Stern, R. M. (2016). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, *24*(7), 1315-1329. https://doi.org/10.1109/TASLP.2016.2545928

Kim, B., Chang, S., Lee, J., & Sung, D. (2021). Broadcasted residual learning for efficient keyword spotting. *arXiv preprint arXiv:2106.04140*. https://doi.org/10.48550/arXiv.2106.04140

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., & Frame, S. (2008). Dysarthric speech database for universal access research. In Ninth Annual Conference of the International Speech Communication Association. https://doi.org/10.21437/Interspeech.2008-480

Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Gang, J., ... & Vergyri, D. (2002). Novel Speech Recognition Models for Arabic ": Johns-Hopkins University Summer Research Workshop 2002, Final Report". In Johns-Hopkins University summer research workshop. https://doi.org/10.1109/ICASSP.2003.1198788

Kirchhoff, K., Vergyri, D., Bilmes, J., Duh, K., & Stolcke, A. (2006). Morphology-based language modeling for conversational Arabic speech recognition. *Computer Speech & Language*, *20*(4), 589-608. https://doi.org/10.1016/j.csl.2005.10.001

Koteswararao, Y. V., & Rao, C. R. (2023). Multichannel KHMF for speech separation with enthalpy based DOA and score based CNN (SCNN). *Evolving Systems*, *14*(3), 501-518. https://doi.org/10.1007/s12530-022-09473-x

Krishna, K., Toshniwal, S., & Livescu, K. (2018). Hierarchical multitask learning for ctc-based speech recognition. *arXiv preprint arXiv:1807.06234*. https://doi.org/10.48550/arXiv.1807.06234

Ladefoged, P., & Johnson, K. (2014). *A course in phonetics*. Cengage learning.

Laureys, T., Vandeghinste, V., & Duchateau, J. (2002). A hybrid approach to compounds in LVCSR. In *Seventh International Conference on Spoken Language Processing*. https://doi.org/10.21437/ICSLP.2002-231

Lee, C. H. (Ed.). (2007). *Advances in Chinese spoken language processing*. World Scientific. https://doi.org/10.1142/6192

Lezhenin, I., Bogach, N., & Pyshkin, E. (2019, September). Urban sound classification using long short-term memory neural network. In *2019 federated conference on computer science and information systems (FedCSIS)* (pp. 57-60). IEEE. https://doi.org/10.15439/2019F185

Li, B., Xie, J. Y., & Rudzicz, F. (2019). Representation learning for discovering phonemic tone contours. *arXiv preprint arXiv:1910.08987*. https://doi.org/10.48550/arXiv.1910.08987

Li, B., Sainath, T. N., Narayanan, A., Caroselli, J., Bacchiani, M., Misra, A., ... & Shannon, M. (2017, August). Acoustic Modeling for Google Home. In *Interspeech* (pp. 399-403). https://doi.org/10.21437/Interspeech.2017-234

Li, X., Dalmia, S., Li, J., Lee, M., Littell, P., Yao, J., ... & Metze, F. (2020, May). Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8249-8253). IEEE. https://doi.org/10.48550/arXiv.2002.11800

Lin, J. C. W., Shao, Y., Djenouri, Y., & Yun, U. (2021). ASRNN: A recurrent neural network with an attention model for sequence labeling. *Knowledge-Based Systems*, *212*, 106548. https://doi.org/10.1016/j.knosys.2020.106548

Lin, J., Kilgour, K., Roblek, D., & Sharifi, M. (2020, May). Training keyword spotters with limited and synthesized speech data. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7474-7478). IEEE. https://doi.org/10.48550/arXiv.2002.01322

Lin, S. H., Yeh, Y. M., & Chen, B. (2010). Extractive speech summarization-From the view of decision theory. In *Eleventh Annual Conference of the International Speech Communication Association*. https://doi.org/10.1016/j.ipm.2011.12.002

Liu, C., Zhang, Q., Zhang, X., Singh, K., Saraf, Y., & Zweig, G. (2019). Multilingual graphemic hybrid ASR with massive data augmentation. *arXiv preprint arXiv:1909.06522*. https://doi.org/10.48550/arXiv.1909.06522

Liu, S., Geng, M., Hu, S., Xie, X., Cui, M., Yu, J., ... & Meng, H. (2021). Recent progress in the CUHK dysarthric speech recognition system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 2267-2281. https://doi.org/10.1109/TASLP.2021.3091805

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., ... & Guo, B. (2022). Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12009-12019). https://doi.org/10.48550/arXiv.2111.09883

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022). https://doi.org/10.48550/arXiv.2103.14030

Loots, L., & Niesler, T. (2011). Automatic conversion between pronunciations of different English accents. *Speech Communication*, *53*(1), 75-84. https://doi.org/10.1016/j.specom.2010.07.006

Lu, Y., Huang, M., Li, H., Guo, J., & Qian, Y. (2020, October). Bi-encoder transformer network for mandarin-english code-switching speech recognition using mixture of experts. In *Interspeech* (pp. 4766-4770). https://doi.org/10.21437/Interspeech.2020-2485

Lüdeling, A., & Kytö, M. (Eds.). (2008). *Corpus Linguistics. Volume 1*. Walter de Gruyter. https://doi.org/10.1515/booksetHSK29

Luo, Y., Chen, Z., & Yoshioka, T. (2020, May). Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 46-50). IEEE. https://doi.org/10.48550/arXiv.1910.06379

Luo, Y., & Mesgarani, N. (2018, April). Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 696-700). IEEE. https://doi.org/10.1109/TASLP.2019.2915167

Luo, Yi, and Nima Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation." *IEEE/ACM transactions on audio, speech, and language processing* 27, no. 8 (2019): 1256-1266. https://doi.org/10.1109/TASLP.2019.2915167

Ma, B., Guan, C., Li, H., & Lee, C. H. (2002, September). Multilingual speech recognition with language identification. In *INTERSPEECH*. https://doi.org/10.21437/ICSLP.2002-178
Majumdar, S., & Ginsburg, B. (2020). Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition. *arXiv preprint arXiv:2004.08531*. https://doi.org/10.48550/arXiv.2004.08531

Malcangi, M., & Grew, P. (2017). Evolving connectionist method for adaptive audiovisual speech recognition. *Evolving Systems*, *8*, 85-94. https://doi.org/10.1007/s12530-016-9156-6

Mazumder, M., Chitlangia, S., Banbury, C., Kang, Y., Ciro, J. M., Achorn, K., ... & Reddi, V. J. (2021, August). Multilingual spoken words corpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Mehra, S., & Susan, S. (2021). Improving word recognition in speech transcriptions by decision-level fusion of stemming and two-way phoneme pruning. In *Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10* (pp. 256-266). Springer Singapore. https://doi.org/10.1007/978-981-16-0401-0_19

Mehra, S., & Susan, S. (2022, December). Early Fusion of Phone Embeddings for Recognition of Low-Resourced Accented Speech. In *2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST)* (pp. 1-5). IEEE. https://doi.org/10.1109/AIST55798.2022.10064735

Mehra, S., & Susan, S. (2023). Deep fusion framework for speech command recognition using acoustic and linguistic features. *Multimedia Tools and Applications*, 1-25. https://doi.org/10.1007/s11042-023-15118-1

Mehra, S., Ranga, V., & Agarwal, R. (2023). Improving speech command recognition through decision-level fusion of deep filtered speech cues. Signal, Image and Video Processing, 1-9. https://doi.org/10.1007/s11760-023-02845-z

Metze, F., Anguera, X., Barnard, E., Davel, M., & Gravier, G. (2013, May). The spoken web search task at MediaEval 2012. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8121-8125). IEEE. https://doi.org/10.1109/ICASSP.2013.6639247

Metze, F., Ding, D., Younessian, E., & Hauptmann, A. (2013). Beyond audio and video retrieval: topic-oriented multimedia summarization. *International Journal of Multimedia Information Retrieval*, *2*, 131-144. https://doi.org/10.1007/s13735-012-0028-y

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*. https://doi.org/10.48550/arXiv.1712.09405

Morshed, M. M., & Ahsan, A. O. (2021). Attention-free keyword spotting. *arXiv preprint arXiv:2110.07749*. https://doi.org/10.48550/arXiv.2110.07749

Mukhamadiyev, A., Khujayarov, I., Djuraev, O., & Cho, J. (2022). Automatic speech recognition method based on deep learning approaches for Uzbek language. *Sensors*, *22*(10), 3683. https://doi.org/10.3390/s22103683

Nahar, K. M., Abu Shquier, M., Al-Khatib, W. G., Al-Muhtaseb, H., & Elshafei, M. (2016). Arabic phonemes recognition using hybrid LVQ/HMM model for continuous speech recognition. *International Journal of Speech Technology*, *19*, 495-508. https://doi.org/10.1007/s10772-016-9337-5

Narayanan, A., & Wang, D. (2013, May). Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7092-7096). IEEE. https://doi.org/10.1109/ICASSP.2013.6639038

Nasersharif, B., Ebrahimpour, M., & Naderi, N. (2023). Multi-layer maximum mean discrepancy in auto-encoders for cross-corpus speech emotion recognition. *The Journal of Supercomputing*, 1-19. https://doi.org/10.1007/s11227-023-05161-y

Newman, D. (2002). The phonetic status of Arabic within the world's languages: the uniqueness of the lughat al-daad. *Antwerp papers in linguistics.*, *100*, 65-75.

Ng, D., Chen, Y., Tian, B., Fu, Q., & Chng, E. S. (2022, May). Convmixer: Feature interactive convolution with curriculum learning for small footprint and noisy far-field keyword spotting. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3603-3607). IEEE. https://doi.org/10.48550/arXiv.2201.05863

Nga, C. H., Li, C. T., Li, Y. H., & Wang, J. C. (2021, December). A Survey of Vietnamese Automatic Speech Recognition. In *2021 9th International Conference on Orange Technology (ICOT)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICOT54518.2021.9680652

Nguyen, Q. T., & Bui, T. D. (2016). Speech classification using SIFT features on spectrogram images. *Vietnam journal of computer science*, *3*, 247-257. https://doi.org/10.1007/s40595-016-0071-3

Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., & Kashino, K. (2023, June). Masked Modeling Duo: Learning Representations by Encouraging Both Networks to Model the Input. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE. https://doi.org/10.48550/arXiv.2210.14648

O'Neill, P. K., Lavrukhin, V., Majumdar, S., Noroozi, V., Zhang, Y., Kuchaiev, O., ... & Kucsko, G. (2021). Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv preprint arXiv:2104.02014*. https://doi.org/10.48550/arXiv.2104.02014

O'Shaughnessy, D. (2008). Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, *41*(10), 2965-2979. https://doi.org/10.1016/j.patcog.2008.05.008

Ogawa, A., & Hori, T. (2017). Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks. *Speech Communication*, *89*, 70-83. https://doi.org/10.1016/j.specom.2017.02.009

Oh, D., Park, J. S., Kim, J. H., & Jang, G. J. (2021). Hierarchical phoneme classification for improved speech recognition. *Applied Sciences*, *11*(1), 428. https://doi.org/10.3390/app11010428

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5206-5210). IEEE. https://doi.org/10.1109/ICASSP.2015.7178964

Papi, S., Trentin, E., Gretter, R., Matassoni, M., & Falavigna, D. (2021). Mixtures of deep neural experts for automated speech scoring. *arXiv preprint arXiv:2106.12475*. https://doi.org/10.48550/arXiv.2106.12475

Passricha, V., & Aggarwal, R. K. (2019). A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition. *Journal of Intelligent Systems*, *29*(1), 1261-1274. https://doi.org/10.1515/jisys-2018-0372

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543). https://doi.org/10.3115/v1/D14-1162

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, *24*(2), 175-184. https://doi.org/10.1121/1.1906875

Plahl, C., Schlüter, R., & Ney, H. (2010). Hierarchical bottle neck features for LVCSR. In *Eleventh annual conference of the international speech communication association*. https://doi.org/10.21437/Interspeech.2010-375

Porter M (1999) Porter stemming algorithm.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130-137. https://doi.org/10.1108/eb046814

Pratap, V., Sriram, A., Tomasello, P., Hannun, A., Liptchinsky, V., Synnaeve, G., & Collobert, R. (2020). Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters. *arXiv preprint arXiv:2007.03001*. https://doi.org/10.48550/arXiv.2007.03001

Principi, E., Squartini, S., Bonfigli, R., Ferroni, G., & Piazza, F. (2015). An integrated system for voice command recognition and emergency detection based on audio signals. *Expert Systems with Applications*, *42*(13), 5668-5683. https://doi.org/10.1016/j.eswa.2015.02.036

Rabiner, L., & Juang, B. H. (1993). Fundamentals of speech processing.

Rabiner, L. R., & Schafer, R. W. (2007). Introduction to digital speech processing. *Foundations and Trends® in Signal Processing*, *1*(1–2), 1-194. https://doi.org/10.1561/2000000001

Rao, W., Zhang, J., & Wu, J. (2020, March). Improved blstm rnn based accent speech recognition using multi-task learning and accent embeddings. In *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing* (pp. 1-6). https://doi.org/10.1145/3388818.3389159

Ravuri, S., & Stolcke, A. (2015). Recurrent neural network and LSTM models for lexical utterance classification. In *Sixteenth annual conference of the international speech communication association*. https://doi.org/10.21437/Interspeech.2015-42

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*. https://doi.org/10.48550/arXiv.1908.10084

Rayson, S. J., Hachamovitch, D. J., Kwatinetz, A. L., & Hirsch, S. M. (1998). *U.S. Patent No. 5,761,689*. Washington, DC: U.S. Patent and Trademark Office.

Richards, J. C., & Schmidt, R. W. (2013). *Longman dictionary of language teaching and applied linguistics*. Routledge. https://doi.org/10.4324/9781315833835

Sadhu, S., Li, R., & Hermansky, H. (2019, May). M-vectors: sub-band based energy modulation features for multi-stream automatic speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6545-6549). IEEE. https://doi.org/10.1109/ICASSP.2019.8682710

Sadhu, S., & Hermansky, H. (2023, June). Importance of Different Temporal Modulations of Speech: a Tale of two Perspectives. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE. https://doi.org/10.48550/arXiv.2204.00065

Sainath, T. N., Chung, I. H., Ramabhadran, B., Picheny, M., Gunnels, J., Kingsbury, B., ... & Chaudhari, U. (2014). Parallel deep neural network training for lvcsr tasks using blue gene/q. In *Fifteenth Annual Conference of the International Speech Communication Association*. https://doi.org/10.21437/Interspeech.2014-272

Sainath, T. N., Kingsbury, B., & Ramabhadran, B. (2012, March). Auto-encoder bottleneck features using deep belief networks. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4153-4156). IEEE. https://doi.org/10.1109/ICASSP.2012.6288833

Sakashita, Y., & Aono, M. (2018). Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions. *Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*.

Sankari, V. M., Snekhalatha, U., Murugappan, M., Chowdhury, M. E., & Chamkha, Z. A. (2023). Artificial Intelligence-Based Hearing Loss Detection Using Acoustic Threshold and Speech Perception Level. *Arabian Journal for Science and Engineering*, 1-17. https://doi.org/10.1007/s13369-023-07927-1

Schwartz, R., & Makhoul, J. (1975). Where the phonemes are: Dealing with ambiguity in acoustic-phonetic recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *23*(1), 50-53. https://doi.org/10.1109/TASSP.1975.1162629

Seide, F., Fu, H., Droppo, J., Li, G., & Yu, D. (2014). 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth annual conference of*

the international speech communication association. https://doi.org/10.21437/Interspeech.2014-274

Selouani, S. A., Sidi Yakoub, M., & O'Shaughnessy, D. (2009). Alternative speech communication system for persons with severe speech disorders. *EURASIP Journal on Advances in Signal Processing*, *2009*, 1-12. https://doi.org/10.1155/2009/540409

Seo, D., Oh, H. S., & Jung, Y. (2021). Wav2kws: Transfer learning from speech representations for keyword spotting. *IEEE Access*, *9*, 80682-80691. https://doi.org/10.1109/ACCESS.2021.3078715

Seide, F., Li, G., Chen, X., & Yu, D. (2011, December). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 24-29). IEEE. https://doi.org/10.1109/ASRU.2011.6163899

Sercu, T., Saon, G., Cui, J., Cui, X., Ramabhadran, B., Kingsbury, B., & Sethy, A. (2017, March). Network architectures for multilingual speech representation learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5295-5299). IEEE. https://doi.org/10.1109/ICASSP.2017.7953167

Shain, C., & Elsner, M. (2019, June). Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 69-85). https://doi.org/10.18653/v1/N19-1007

Shahamiri, S. R. (2021). Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*,*29*, 852-861. https://doi.org/10.1109/TNSRE.2021.3076778

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE access*, *7*, 53040-53065. https://doi.org/10.1109/ACCESS.2019.2912200

Shi, X., Yang, H., & Zhou, P. (2016, October). Robust speaker recognition based on improved GFCC. In *2016 2nd IEEE international conference on computer and communications (ICCC)* (pp. 1927-1931). IEEE. https://doi.org/10.1109/CompComm.2016.7925037

Shuai, Y., Pengyu, L., Gaojie, W., Dou, F., Yong, T., & Yuwei, S. (2021, August). Mechanical Faults Diagnosis of HVCB Based on Auditory Features and Sparse Representation Classification. In *2021 Power System and Green Energy Conference (PSGEC)* (pp. 650-654). IEEE. https://doi.org/10.1109/PSGEC51302.2021.9542694

Silva, T. C., & Yehia, H. C. (2011). Sonoridade em Artes, Saúde e Tecnologia. *Revista Docência do Ensino Superior*, *1*, 62-74. https://doi.org/10.35699/2237-5864.2011.2021

Sönmez, Y. Ü., & Varol, A. (2020). A speech emotion recognition model based on multi-level local binary and local ternary patterns. *IEEE Access*, *8*, 190784-190796. https://doi.org/10.1109/ACCESS.2020.3031763

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958. https://doi.org/10.5555/2627435.2670313

Stern, R. M., Gouvêa, E., Kim, C., Kumar, K., & Park, H. M. (2008, May). Binaural and multiple-microphone signal processing motivated by auditory perception. In *2008 Hands-Free Speech Communication and Microphone Arrays* (pp. 98-103). IEEE. https://doi.org/10.1109/HSCMA.2008.4538697

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021, June). Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 21-25). IEEE. https://doi.org/10.48550/arXiv.2010.13154

Takashima, R., Takiguchi, T., & Ariki, Y. (2020, May). Two-step acoustic model adaptation for dysarthric speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6104-6108). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9053725

Tan, X., Qin, T., Soong, F., & Liu, T. Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*. https://doi.org/10.48550/arXiv.2106.15561

Theodor, U., Shaked, U., & de Souza, C. E. (1994). A game theory approach to robust discrete-time H/sub/spl infin//-estimation. *IEEE Transactions on Signal Processing*, *42*(6), 1486-1495. https://doi.org/10.1109/78.286964

Tong, Fuchuan, Tao Li, Dexin Liao, Shipeng Xia, Song Li, Qingyang Hong, and Lin Li. "The XMUSPEECH system for accented English automatic speech recognition." *Applied Sciences* 12, no. 3 (2022): 1478. https://doi.org/10.3390/app12031478

Torfi, A., Iranmanesh, S. M., Nasrabadi, N., & Dawson, J. (2017). 3d convolutional neural networks for cross audio-visual matching recognition. *IEEE Access*, *5*, 22081-22091. https://doi.org/10.48550/arXiv.1706.05739

Trinh, V. A., & Braun, S. (2022, May). Unsupervised speech enhancement with speech recognition embedding and disentanglement losses. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 391-395). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9746973

Tu, T., Chen, Y. J., Yeh, C. C., & Lee, H. Y. (2019). End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. *arXiv preprint arXiv:1904.06508*. https://doi.org/10.48550/arXiv.1904.06508

Tur, G., & De Mori, R. (2011). *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons. https://doi.org/10.1002/9781119992691

Turrisi, R., Braccia, A., Emanuele, M., Giulietti, S., Pugliatti, M., Sensi, M., ... & Badino, L. (2021). EasyCall corpus: a dysarthric speech dataset. *arXiv preprint arXiv:2104.02542*. https://doi.org/10.21437/Interspeech.2021-549

Tüske, Z., Schlüter, R., & Ney, H. (2013, August). Multilingual hierarchical MRASTA features for ASR. In *Interspeech* (pp. 2222-2226). https://doi.org/10.1109/ICASSP.2014.6855129

Uebel, L. F., & Woodland, P. C. (1999). An investigation into vocal tract length normalisation. In *Sixth European Conference on Speech Communication and Technology*.

Vachhani, B., Bhat, C., & Kopparapu, S. K. (2018, September). Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. In *Interspeech* (pp. 471-475). https://doi.org/10.21437/Interspeech.2018-1751

Van Huy, N., Mai, L. C., & Thang, V. T. (2014). Vietnamese recognition using tonal phoneme based on multi space distribution. *Journal of Computer Science and Cybernetics*, *30*(1), 28-38. https://doi.org/10.15625/1813-9663/30/1/3553

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. https://dl.acm.org/doi/10.5555/3295222.3295349

Vygon, R., & Mikhaylovskiy, N. (2021). Learning efficient representations for keyword spotting with triplet loss. In *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23* (pp. 773-785). Springer International Publishing. https://doi.org/10.1007/978-3-030-87802-3_69

Vijayalakshmi, P., & Reddy, M. R. (2006, August). Assessment of dysarthric speech and an analysis on velopharyngeal incompetence. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 3759-3762). IEEE. https://doi.org/10.1109/IEMBS.2006.259334

Virtanen, T., Singh, R., & Raj, B. (Eds.). (2012). *Techniques for noise robustness in automatic speech recognition*. John Wiley & Sons. https://doi.org/10.1002/9781118392683

Wang, Z. Q., Wang, P., & Wang, D. (2020). Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR. *IEEE/ACM transactions on audio, speech, and language processing*, *28*, 1778-1787. https://doi.org/10.1109/TASLP.2020.2998279

Watanabe, S., Hori, T., & Hershey, J. R. (2017, December). Language independent end-to-end architecture for joint language identification and speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 265-271). IEEE. https://doi.org/10.1109/ASRU.2017.8268945

Wazir, A. S. M. B., & Chuah, J. H. (2019, June). Spoken Arabic digits recognition using deep learning. In *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)* (pp. 339-344). IEEE. https://doi.org/10.1109/I2CACIS.2019.8825004

Wei, Y., Gong, Z., Yang, S., Ye, K., & Wen, Y. (2022). EdgeCRNN: an edge-computing oriented model of acoustic feature enhancement for keyword spotting. *Journal of Ambient Intelligence and Humanized Computing*, 1-11. https://doi.org/10.1007/s12652-021-03022-1

Weide, R. (1998). The carnegie mellon pronouncing dictionary. *release 0.6, www. cs. cmu. edu*.

Wells, D., & Richmond, K. (2021, August). Cross-lingual transfer of phonological features for low-resource speech synthesis. In *Proc. 11th ISCA Speech Synth. Workshop* (pp. 160-165). https://doi.org/10.21437/SSW.2021-28

Whitehill, T. L., & Ciocca, V. (2000). Speech errors in Cantonese speaking adults with cerebral palsy. *Clinical linguistics & phonetics*, *14*(2), 111-130. https://doi.org/10.1080/026992000298869

Winata, G. I., Lin, Z., & Fung, P. (2019, August). Learning multilingual meta-embeddings for code-switching named entity recognition. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)* (pp. 181-186). https://doi.org/10.18653/v1/W19-4320

Woszczyk, D., Petridis, S., & Millard, D. (2020). Domain adversarial neural networks for dysarthric speech recognition. *arXiv preprint arXiv:2010.03623*. https://doi.org/10.48550/arXiv.2010.03623

Wu, J., Hong, R., & Tian, Q. (2022). Special issue on cross-modal retrieval and analysis. *International Journal of Multimedia Information Retrieval*, *11*(4), 523-524. https://doi.org/10.1007/s13735-022-00265-2

Xie, X., Ruzi, R., Liu, X., & Wang, L. (2022). Variational auto-encoder based variability encoding for dysarthric speech recognition. *arXiv preprint arXiv:2201.09422*. https://doi.org/10.48550/arXiv.2201.09422

Xie, X., Sui, X., Liu, X., & Wang, L. (2022). Investigation of Deep Neural Network Acoustic Modelling Approaches for Low Resource Accented Mandarin Speech Recognition. *arXiv preprint arXiv:2201.09432*. https://doi.org/10.48550/arXiv.2201.09432

Xiong, F., Barker, J., & Christensen, H. (2018, October). Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition. In *Speech Communication; 13th ITG-Symposium* (pp. 1-5). VDE.

Xiong, F., Barker, J., & Christensen, H. (2019, May). Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5836-5840). IEEE. https://doi.org/10.1109/ICASSP.2019.8683091

Xu, Y., Du, J., Dai, L. R., & Lee, C. H. (2014). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(1), 7-19. https://doi.org/10.1109/TASLP.2014.2364452

Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., ... & Chen, X. (2019, May). LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)* (pp. 1-8). IEEE. https://doi.org/10.48550/arXiv.1810.06990

Ye, Z., Hu, S., Li, J., Xie, X., Geng, M., Yu, J., ... & Meng, H. (2021, June). Development of the cuhk elderly speech recognition system for neurocognitive disorder detection using the dementiabank corpus. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6433-6437). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9413634

Yenkimaleki, M., & van Heuven, V. J. (2021). Effects of attention to segmental vs. suprasegmental features on the speech intelligibility and comprehensibility of the EFL learners targeting the perception or production-focused practice. *System*, *100*, 102557. https://doi.org/10.1016/j.system.2021.102557

Yeung, A. H. L. (2020). Revisiting phonotactic repairs in Cantonese loanword phonology: it's all about sC. *Journal of East Asian Linguistics*, *29*, 279-309. https://doi.org/10.1007/s10831-020-09212-w

Yi, J., Tao, J., Wen, Z., & Bai, Y. (2018, April). Adversarial multilingual training for low-resource speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4899-4903). IEEE. https://doi.org/10.1109/ICASSP.2018.8461771

Yu, J., Xie, X., Liu, S., Hu, S., Lam, M. W., Wu, X., ... & Meng, H. (2018, September). Development of the CUHK Dysarthric Speech Recognition System for the UA Speech Corpus. In *Interspeech* (pp. 2938-2942). https://doi.org/10.21437/Interspeech.2018-1541

Zhang, Q., Zhang, H., Zhou, K., & Zhang, L. (2023). Developing a Physiological Signal-Based, Mean Threshold and Decision-Level Fusion Algorithm (PMD) for Emotion Recognition. *Tsinghua Science and Technology*, *28*(4), 673-685. https://doi.org/10.26599/TST.2022.9010038

Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C. C., Pang, R., ... & Wu, Y. (2020). Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*. https://doi.org/10.48550/arXiv.2010.10504

Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., & Gutierrez-Osuna, R. (2018, September). L2-ARCTIC: A non-native English speech corpus. In *Interspeech* (pp. 2783-2787). https://doi.org/10.21437/Interspeech.2018-1110

Zhu, J., Huang, C., & De Meo, P. (2023). DFMKE: A dual fusion multi-modal knowledge graph embedding framework for entity alignment. *Information Fusion*, *90*, 111-119. https://doi.org/10.1016/j.inffus.2022.09.012

Zhu, Y., Haghani, P., Tripathi, A., Ramabhadran, B., Farris, B., Xu, H., ... & Zhang, Q. (2020). Multilingual Speech Recognition with Self-Attention Structured Parameterization. In *INTERSPEECH* (pp. 4741-4745). https://doi.org/10.21437/interspeech.2020-2847

Zia, T., & Zahid, U. (2019). Long short-term memory recurrent neural network architectures for Urdu acoustic modeling. *International Journal of Speech Technology*, *22*, 21-30. https://doi.org/10.1007/s10772-018-09573-7

Zeng, M., & Xiao, N. (2019). Effective combination of DenseNet and BiLSTM for keyword spotting. *IEEE Access*, *7*, 10767-10775. https://doi.org/10.1109/ACCESS.2019.2891838

BRIEF BIO-DATA

Name: Ms. Sunakshi Mehra
Registration Number: 2K19/PHDIT/02
Department of Information Technology
Delhi Technological University, Delhi - 110042
Email: mehra.sunakshi623@gmail.com
ResearchGate Profile:
https://www.researchgate.net/profile/Sunakshi-Mehra-2
Google Scholar:
https://scholar.google.com/citations?user=kpJ71HEAAAAJ&hl=en
ORCID iD: 0000-0002-6397-6049

Educational Qualification: B. Tech in Information Technology from Guru Gobind Singh Indraprastha University, Delhi in 2015 with 70.27 grade.
M. Tech. in Information Technology from Guru Gobind Singh Indraprastha University, Delhi in 2017 with 8.222 grade.
Joined Ph.D. in Information Technology Department at Delhi Technology University Delhi in July 2019.
Cleared UGC NET examination.
Research areas of interest: Speech and Natural language Processing, deep learning, linguistics.