SPEAKNet: Spectrogram-Phoneme Embedding Architecture for Knowledge-enhanced Speech Command Recognition

Sunakshi Mehra
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India

ABSTRACT

This research aims to enhance automatic speech recognition (ASR) by integrating multimodal data—specifically, text transcripts and Mel spectrograms generated from raw audio signals. The study explores the often-overlooked role of phonological features and spectrogram-based representations in improving the accuracy of spoken word recognition. A dual-path approach is adopted: EfficientNetV2 is utilized to extract features from spectrogram images, while a Speech2Text transformer model is employed to generate text transcripts. For evaluation, the study uses ten-word categories from version 2 of the Google Speech Commands dataset. To reduce noise in the audio samples, a Kalman filter is applied, ensuring cleaner signal processing. The resulting Mel spectrograms are resized to 256×256 pixels to produce two-dimensional visual representations of the audio data. These images are then classified using EfficientNetV2, pre-trained on the ImageNet dataset. In parallel, a grapheme-to-phoneme (G2P) model is used to convert Speech2Text outputs into phonemes. These are further processed through a technique called phoneme slicing, which extracts core phonological units—such as fricatives, nasals, liquids, glides, plosives, approximants, taps/flaps, trills, and vowels—based on articulatory features like manner and place of articulation. The proposed system employs a late fusion strategy that combines phoneme embeddings with image-based embeddings to achieve high classification accuracy. This fusion not only boosts ASR performance but also underscores the value of incorporating linguistic and phonological knowledge into spoken language understanding. Through comprehensive ablation analysis, the study demonstrates that the integration of spectrograms and phonological analysis sets a new benchmark, outperforming existing models in terms of accuracy and interpretability.

General Terms

Pattern Recognition

Keywords

Speech Command Recognition, EfficientNetV2, Speech Filtering Techniques, Transformer Models, Feedforward Neural Network (FNN), Multimodal Speech Processing, Mel Spectrogram, Phoneme Analysis

1. INTRODUCTION

The world we live in is full of rich sensory information—what we see, hear, and read all come together to shape how we perceive and understand our surroundings [1, 2]. In the field of speech technology, Audio-Visual Speech Recognition (AVSR) taps into both sound and visual cues, such as lip movements, to interpret spoken language. This combined approach is especially useful in noisy environments, where lip movements provide reliable clues even when audio signals are unclear [3]. Building such systems requires not just the raw speech signals but also an understanding of context and language structure. ASR, a key AI technology, enables machines to interpret and respond to spoken language. Typically, ASR systems work with audio inputs (like .wav files), which go through processes like noise reduction and feature extraction—often involving spectrograms. From there, an acoustic model identifies phonemes, and Hidden Markov Models (HMMs) are used to predict the most likely sequence of words based on a language model. Accurate ASR performance depends heavily on clear pronunciation. Mispronunciations or accent variations, especially across speakers from different linguistic backgrounds, can reduce recognition accuracy. This becomes particularly important in contexts like English as a Foreign Language (EFL) education, where ASR systems can help learners improve their pronunciation and fluency. In such cases, detecting and correcting pronunciation errors becomes a crucial part of making these tools more effective and supportive for learners [4].

Speech-related technologies generally follow one of two approaches: a knowledge-driven approach, where experts provide standardized or canonical speech samples, and a data-driven approach, which leverages large collections of varied, non-standard speech recordings. While knowledge-driven methods are typically more effective for generating speech, data-driven approaches tend to perform better in recognizing speech, identifying pronunciation errors, and managing the wide range of pronunciation variations found in real-world speech. This flexibility makes data-driven methods particularly valuable for handling diverse speaker populations. In the broader field of natural language processing (NLP), data-driven strategies—especially those based on supervised machine learning—have become essential for tackling complex language-related tasks. These include optical character recognition (OCR), document classification, and sentiment analysis, where models learn patterns from annotated datasets to make accurate predictions [5, 6].

Understanding phonological errors is crucial, as these errors are often shaped by a learner's native language and tend to be specific to each linguistic background [7, 8]. For instance, the phonological features of a learner's first language can significantly influence how English is pronounced, often leading to systematic variations and errors. Addressing this issue, the present study focuses on the phonological challenges encountered by Arab learners of EFL—a group that has received limited attention in existing research. To bridge this gap, the study proposes an automated system designed to detect and correct phonological errors in English speech. The goal is to provide learners with timely, personalized feedback, thereby supporting more effective pronunciation practice and fostering greater motivation to improve their spoken English skills.

Recent progress in large-scale ASR architectures has demonstrated significant improvements in English speech recognition tasks [9, 10]. In particular, models trained with self-supervised learning objectives—such as wav2vec 2.0 [11], w2v-BERT [12], and BigSSL [13]—have further pushed the boundaries of performance. These advances build upon traditional supervised ASR frameworks by incorporating vast amounts of annotated data. However, the effectiveness of such models remains heavily dependent on the availability of large-scale training datasets. It is important to note that simply increasing model size does not necessarily yield better performance, especially in contexts where training data is limited—such as non-English languages or other low-resource scenarios. Since much of the current ASR research and datasets are centered around English [13], there is an ongoing challenge in adapting these high-performing English ASR models—such as RNN-T [14]—to support other languages effectively [15]. Addressing this challenge holds the potential to extend the benefits of ASR to a broader, more linguistically diverse global population. The main contributions of this paper are as follows:

• The research enhances ASR by merging text transcripts and mel spectrograms, emphasizing the unexplored potential of spectrograms and phonology in improving spoken word accuracy.

• Employing a dual method involving the Speech2Text transformer and the EfficientNetV2, this study utilizes the Google Speech Command dataset version 2 with ten-word categories. Mel spectrogram images are resized to 256 x 256 pixels and classified using ImageNet and EfficientNetV2.

• Phoneme slicing, which extracts vital phonological elements while considering articulation, is integrated. A late fusion method, blending phone and image embeddings, delivers exceptional accuracy and underscores the significance of phonological analysis in speech interpretation, establishing a new benchmark.

The paper is structured into the following sections: Section 2 offers a comprehensive review of speech recognition literature, Section 3 introduces the proposed dense architecture for consolidating posterior scores, Section 4 delineates the experimental framework and summarizes the outcomes, and finally, Section 5 concludes the paper and suggests avenues for future research.

2. RELATED WORK

Working with spoken language data, whether for developing language technologies or conducting (socio)linguistic research, often requires high-quality transcriptions. These transcripts are usually created at the orthographic or word level, but depending on the specific application, additional layers of annotation—such as sentiment, syntactic structure, or phonetic features—may also be included. However, manual transcription remains a time-consuming and resource-intensive task, posing a significant bottleneck for many projects. To address this challenge, there is increasing interest in using ASR as a tool within the transcription and annotation workflow. This typically involves generating automated transcripts using ASR, which are later reviewed and refined by human annotators. For example, DARLA [16], a tool developed for linguistic research, employs ASR through its BedWord service to generate fully automated transcripts for audio data. Similarly, [17] examine the role of ASR in supporting under-resourced languages, contributing to their documentation and preservation. In another example, [18] demonstrate how ASR can be effectively integrated into language documentation workflows, offering tangible benefits for linguists in the field. Further, [19] present an end-to-end ASR model that not only generates orthographic transcripts but also provides additional linguistic annotations, such as phoneme sequences and part-of-speech tags.

This work builds upon an autoregressive Transformer network designed to support multiple speech-related tasks, including ASR, speech translation, and speech synthesis. In earlier research, distinct model architectures were typically developed for each individual task in the speech processing domain. For ASR, widely adopted frameworks include Connectionist Temporal Classification (CTC) [20, 21], Attention-based Encoder-Decoder (AED) networks [22], and Transducer models that incorporate either Recurrent Neural Networks (RNNs) or Transformer-based architectures [23]. In the area of speech synthesis, AED-based methods such as Tacotron [24], Tacotron 2 [25], and TransformerTTS [26] have become popular due to their ability to generate high-quality speech. Additionally, duration-based approaches like FastSpeech [27], FastSpeech 2 [28], and RobuTrans [29] have been widely used for their efficiency and robustness in generating natural-sounding speech. Beyond speech tasks, a variety of Transformer-based architectures—including encoder-only models [30], decoder-only models [31], and encoder-decoder configurations [32, 33]—have been extensively applied in broader NLP applications. These include tasks such as machine translation, text summarization, and question answering, reflecting the versatility and effectiveness of Transformer networks across modalities.

Self-supervised pre-trained models have brought significant advancements in the field of low-resource speech recognition. By leveraging large volumes of unlabeled multilingual speech data, these models are capable of learning cross-lingual phoneme-level representations, enabling them to generalize across various languages—even those with limited labeled resources. Fine-tuning these multilingual pre-trained models has shown promising results, often achieving notably low Word Error Rates (WER), even when only a small amount of task-specific data is available [34, 35]. However, deploying such powerful models in real-world scenarios, especially on resource-constrained devices like smartphones and laptops, poses a significant challenge. Models such as XLS-R and XLSR-53, while highly effective, consist of hundreds of millions of parameters, which makes them unsuitable for devices with limited computational power and memory. To bridge this gap between model performance and practical deployment, there is a growing need for lightweight multilingual speech models—particularly for use in industrial applications and minority language contexts. One promising solution is model pruning, a technique designed to reduce the number of parameters in a neural network without significantly affecting its accuracy. Guided by the lottery ticket hypothesis, researchers have shown that it is possible to extract a smaller, efficient subnetwork from a larger model that can perform comparably well [36]. A notable approach in this direction is PARP (Pruning and Auto-Regressive Fine-Tuning), which introduces a systematic method to identify and fine-tune these sparse subnetworks within self-supervised speech representation models. Through iterative pruning and fine-tuning, PARP effectively reduces model complexity while preserving performance, making it a valuable technique for bringing state-of-the-art speech recognition to low-resource and mobile environments [37].

Deep neural networks (DNNs) have brought transformative improvements across a range of fields, including computer vision (CV) [38] and NLP [39]. In the domain of speech processing, DNNs have demonstrated clear advantages over traditional methods, particularly in tasks such as phoneme recognition (PR) and ASR [40]. Their strength lies in the ability to learn complex hierarchical patterns from large volumes of labeled data [41]. Despite these advancements, DNN-based models often face limitations when applied in low-resource settings, where annotated data is scarce. This gap has led to growing interest in SSL as an emerging paradigm within deep learning research [42]. SSL offers a promising alternative by enabling models to learn directly from raw input data, without requiring extensive manual annotation. In SSL, models are initially trained on unlabeled data using pretext tasks that allow them to learn meaningful representations. These pretrained models are then fine-tuned on specific downstream tasks, such as phoneme recognition and ASR, using a smaller amount of labeled data. This approach not only enhances model performance in low-resource environments but also broadens the applicability of deep learning in speech-related tasks by reducing dependence on large annotated datasets.

Despite the impressive progress made with SSL models, several challenges persist. One notable limitation is that training these models typically demands large volumes of unlabeled audio data, high computational power—often requiring multiple GPUs—and extended training durations. These resource-intensive requirements can hinder accessibility, particularly for researchers or developers working in low-resource environments. Additionally, there has been limited exploration of strategies for training SSL models effectively when only a small amount of data is available [41]. In the field of self-supervised speech representation learning, existing approaches can broadly be categorized into three main types, each distinguished by its underlying training objective: generative, contrastive, and predictive learning [43, 44, 45]. These frameworks guide the model in learning meaningful patterns from unlabelled data, laying the foundation for improved performance in downstream speech tasks.

3. PROPOSED METHODOLOGY

In this section, we explore decision-level fusion techniques, which have shown promising results in enhancing classification performance [70–74]. Building on the approach proposed by Mehra et al. (2024), we adopt a strategy that utilizes the maximum weighted score to integrate outputs from two distinct channels.

Recent findings by Zhang et al. (2023) further emphasize the effectiveness of late fusion, where features extracted from different segments of an enhanced input signal are combined, leading to improved categorization outcomes. After the training stage, late fusion has been widely adopted by researchers as a method for merging outputs from multiple modalities into a unified representation. In our work, we employ this technique by combining probabilistic scores from different modalities, which is expected to significantly enhance the accuracy of our speech recognition system.

3.1 Text-transcripts pre-processing

During the data preprocessing phase, we begin by generating text transcripts from audio samples using a pre-trained Speech2Text model. These transcripts are then converted into phonemes through a grapheme-to-phoneme (G2P) model. To accurately capture the sounds, stress patterns, and articulatory features of spoken words, we utilize the CMU Pronouncing Dictionary, a comprehensive resource containing phonetic transcriptions for over 125,000 English words. After cleaning and normalizing the text, we extract phonemes from the aligned transcripts, paying particular attention to syllable structures and stress. For consistency, we adopt the standard phonetic representation of each word and omit the numerical stress markers—such as 0 (no stress), 1 (primary stress), and 2 (secondary stress)—to focus purely on the phonemic content.

For example, the word about might be transcribed as [AH0, B, AW1, T], from which we retain only the phoneme sequence [AH, B, AW, T] after removing the stress indicators. Each transcript is segmented into individual words, which are then represented by their corresponding phonemes. It's important to note that exact phoneme matches may vary due to differences in pronunciation across speakers, which adds an additional layer of complexity to the alignment process [48]. The detailed categorization of phonological features is illustrated in Fig. 1.

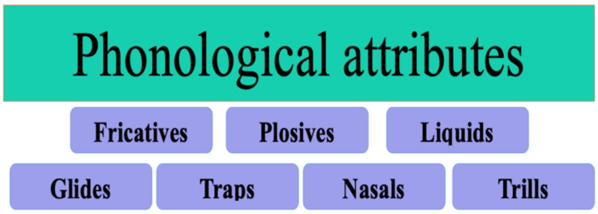

Fig 1: Various types of phonological attributes

3.2 Phoneme transformation and embeddings

The phonemes, along with suprasegmental features such as stress patterns [50], are converted into numerical embeddings using XLNet [49]. XLNet, an advanced variant of the Transformer-XL architecture, is trained using an autoregressive learning objective. Unlike traditional models, XLNet captures bidirectional context by maximizing the likelihood over all possible permutations of the input sequence, allowing it to understand both past and future dependencies more effectively. Each resulting embedding has a dimensionality of 768 × 1, representing the rich phonological and contextual features of the input. These embeddings are then aggregated to generate final representations for each word category. The resulting vectors serve as inputs to the proposed neural network architecture, enabling the model to process and classify spoken words with improved accuracy and contextual awareness.

In this study, specific phonological features are extracted from the phoneme representations, focusing on categories such as fricatives, nasals, liquids, glides, plosives, approximants, taps/flaps, trills, and vowels. These features are analyzed by considering both the manner and place of articulation. To ensure consistency in phonetic representation, the CMU Pronouncing Dictionary is employed, which adheres to American English pronunciation standards and reflects principles of the International Phonetic Alphabet (IPA). The method captures detailed articulatory distinctions. For instance, plosives—also known as stop consonants—briefly block airflow and include voiced sounds like /b/, /d/, /g/ and voiceless ones like /p/, /t/, /k/. Fricatives, such as /f/, /s/, /v/, and /z/, are characterized by their high amplitude and turbulent airflow. Nasals (e.g., /m/, /n/, /ŋ/) allow air to pass through the nasal cavity, while glides (or semivowels) like /j/ and /w/ involve smooth transitions between articulatory positions. Taps and flaps represent rapid, single-contact sounds, such as the American English /t/ in "butter," whereas trills are produced by the vibration of an articulator, commonly found in languages like Spanish. Liquids, including /l/ and /r/, involve minimal constriction and allow continuous airflow, while approximants are produced with articulators that come close without creating turbulence. Vowels, distinguished by their relatively high pitch and amplitude, contrast with consonants in terms of airflow and vocal tract configuration. The analysis accounts for articulation points—such as dorsal, labial, coronal, and radical—providing a comprehensive framework for phoneme classification. Although the CMU dictionary provides a reliable phonetic foundation, variations in pronunciation across speakers pose challenges. To mitigate this, the approach incorporates phoneme filtering and normalization strategies, offering a unified representation of speech sounds.

Once the distinct phoneme classes—namely fricatives, vowels, plosives, nasals, and glides—are isolated, they are transformed into phoneme embeddings using the XLNet model. These embeddings are organized based on their phoneme types and maintain a dimensionality of 768 × 1 for each word category. The resulting phoneme representations are then input into the proposed neural network architecture for further processing. To assess the contribution of each phoneme type and their segmented phonological patterns to spoken word recognition, an ablation analysis is performed. This evaluation helps determine the individual and collective impact of these features on the overall system performance.

3.3 Audio denoising and pre-processing

The processing begins with the application of the Kalman filter to the raw audio samples. In the context of audio processing, the Kalman filter plays a crucial role in distinguishing the desired speech signal from background noise or interference. It operates by modeling the system state—encompassing both the signal and noise components—and iteratively updating its estimates with each new observation. This adaptive estimation process enables effective noise suppression and enhancement of the overall audio quality. Kalman filters are widely used in domains such as speech enhancement, echo cancellation, active noise control, sonar and radar processing, and audio restoration. Following noise reduction, the cleaned audio signals are converted into two-dimensional mel spectrogram images using the Librosa library. These spectrograms undergo several pre-processing transformations, including translation, rotation, and resizing, to standardize the data. All spectrogram images are resized to a fixed dimension of 256 × 256 pixels to ensure consistency across the dataset. These uniformly pre-processed images are then input into the EfficientNetV2 model. The posterior probabilities generated by this model are subsequently passed to the proposed neural architecture illustrated in Fig. 2 for further processing.

3.4 Proposed architecture

The integrated visual and textual embeddings, obtained from earlier stages, are each processed through two separate feedforward neural network architectures tailored to their respective feature modalities. The first model is designed to handle the posterior scores derived from the EfficientNetV2 output. This model begins with a flattening layer and is followed by a sequence of fully connected dense layers with batch normalization. These layers consist of 1024, 512, 256, and 64 units, respectively, and utilize the hyperbolic tangent ("tanh") activation function, which aids in capturing both positive and negative signal representations. In parallel, the second model is structured to process posterior scores obtained from phoneme embeddings—capturing the linguistic information of the spoken data.

This network also starts with a flattening layer, followed by dense layers containing 512, 256, and 64 units. Unlike the first, this model does not apply batch normalization and instead uses the Rectified Linear Unit ("ReLU") activation function, which is well-suited for learning sparse and efficient representations. Both models are trained using the stochastic gradient descent (SGD) optimization strategy, with adaptive learning rate control provided by the Adam optimizer (Duchi et al., 2011) [51], known for its robust performance across a wide range of deep learning applications. The loss function employed for training is sparse categorical cross-entropy, which is particularly effective for multi-class classification tasks involving discrete target labels.

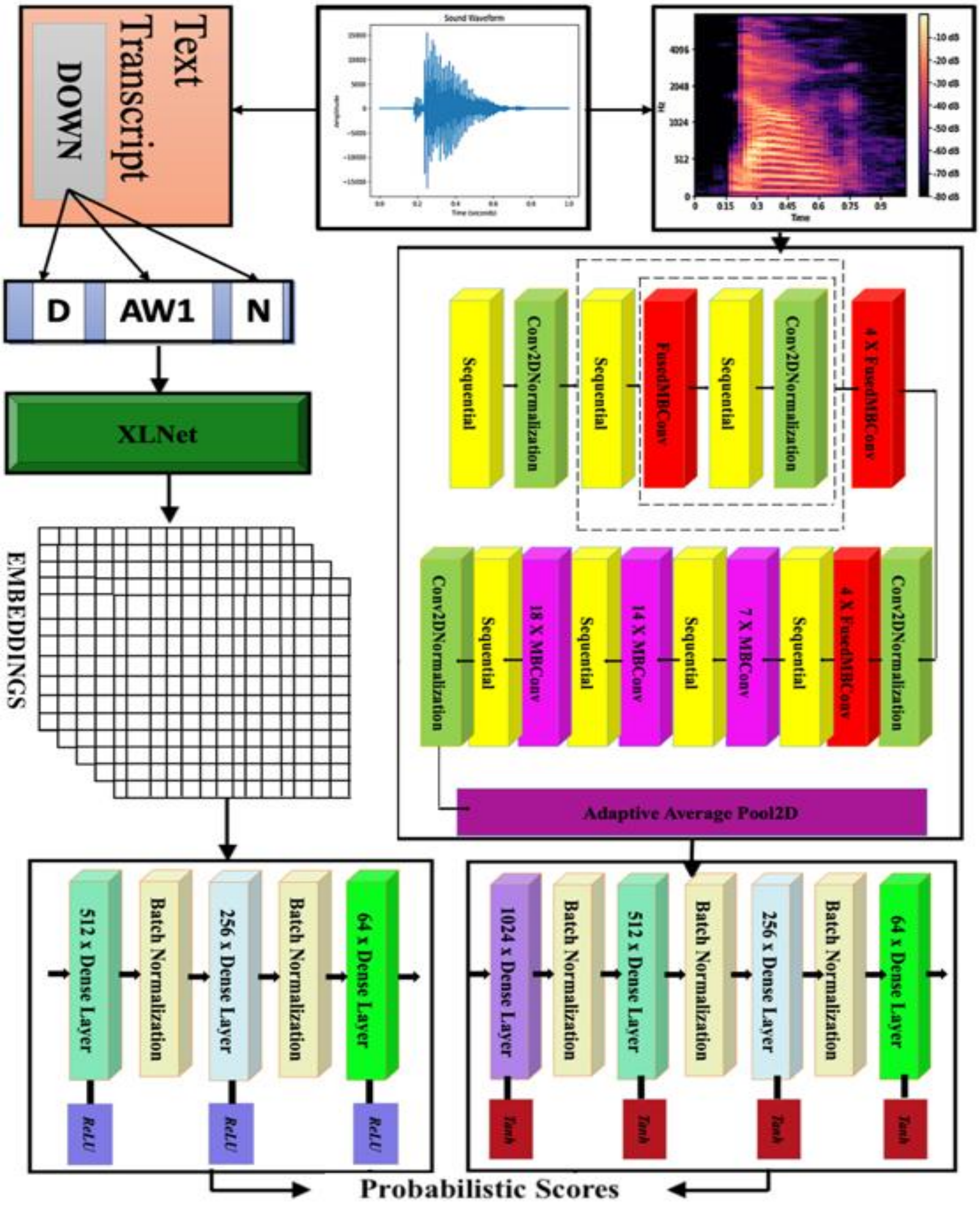

Fig 2: A Multimodal Framework for Enhanced Speech Recognition via Phoneme Embeddings and Spectrogram Features

4. EXPERIMENTAL RESULTS

4.1 Hyperparameters

Our experiments were conducted using Python 3.10.0 on a macOS Big Sur system powered by the Apple M1 chip. To ensure transparency and encourage reproducibility, we have made our code publicly available for future research. Given the computational intensity of our work—particularly in processing audio data and running transformer-based models—we utilized Google Colab Pro with GPU acceleration to manage training and evaluation more efficiently. It’s important to note that Colab Pro offers up to 32 GB of RAM, a limitation we accounted for during model training and memory management. We used the Librosa library extensively for audio processing tasks, which helped streamline feature extraction and spectrogram generation. In our core experiment, we utilized pre-trained probabilistic scores generated by the EfficientNetV2 transformer as input to a custom neural network. The model was trained over 100 epochs, allowing it to effectively learn from the data. We adopted the ReLU activation function to introduce non-linearity, which has proven beneficial for classification tasks. For optimization, we employed the Adam algorithm, chosen for its adaptive learning rate and robust convergence properties. To convert spoken language into text, we used the Speech2Text transformer. These transcripts were then mapped to phonemes and phonological patterns to reflect the actual pronunciation of words. For this step, we leveraged the CMU Pronouncing Dictionary—a widely used resource that provides standardized phonetic representations for English words. This conversion not only improved transcript accuracy but also facilitated detailed linguistic analysis, including the identification of stress markers, articulation types, and sound properties relevant to speech recognition and phonological error detection.

4.2 Ablation analysis

In this section, we present an ablation study aimed at evaluating the contribution of both audio-based and linguistic features, as well as the effectiveness of their combined use, on the performance of our dense neural network model. Through a detailed analysis of the decision-level fusion framework integrated into our model, we systematically examine the role of individual audio and text-based components. This includes assessing their pairwise interactions across different speaker groups (e.g., male vs. female) and analyzing the underlying characteristics of the deep learning architecture itself. The findings from this ablation study provide valuable insights into the strengths and limitations of each feature type and the conditions under which their fusion leads to performance improvements. A summary of the key observations is presented below.

Here are the key technical findings:

- In our dense framework, image-based classification outperforms text-based classification.
- Dual branching audio-visual and text modalities perform better than other individual audio/visual/text components.
- When classifying spoken words, the audio-visual-based outperforms text-based approach.
- A future area of exploration involves utilizing spoken utterances for the classification of unlabeled speech data, which is deemed highly necessary.
- Combining phoneme categories for 10-word subjects achieves test accuracies of 90.20% with stress markers (768 X 1 embeddings).
- Ablation analysis using XLNet-transformer shows varying accuracy for different phoneme combinations as shown in Table 3. Stress markers are crucial for identifying spoken words from text transcripts.
- In summary, dual modalities are effective. However, EfficientNetV2 performed better than XLNet for speech command classification.

Table 1 presents a detailed comparison of our proposed late fusion strategy—which integrates audio-visual and text features through a neural layered model—with several state-of-the-art methods.

Table 1. Assessing performance in comparison to the state-of-the-art results for the 10-word category within the Google Speech Command Dataset

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-XLNet)	**99.80%**

All models were assessed under uniform conditions, utilizing the same dataset and similar experimental settings to ensure a fair comparison. Compared to the approach proposed by Haque et al. (2020), which used a convolutional neural network (CNN) with MFCC features, our method achieved a significant improvement in accuracy. Notably, our model reached a test accuracy of 99.80% across ten spoken word categories, emphasizing its strength and reliability, especially in well-resourced scenarios where ample training data is available.

Table 2. Improving Speech Recognition Using Universal Sentence Encoder on the 10-Word Google Speech Commands Dataset Without Phonological Stress Markers

APPROACH	**Methodology proposed on 10-word categories**	**TRAIN ACCURACY (%)**	**TEST ACCURACY (%)**
Phonetic approach	Phonemes (With stress markers) XLNet	92.40	90.20
Visual-based approach	EfficientNet V2	99.99	99.78

Decision-level Fusion	**Phonemes (With stress markers) XLNet + EfficientNet V2**	_	**99.80**

Table 3. Evaluation metrics to check the effect of stress markers and spectrograms on the 10-word Google Speech Commands dataset

LINGUISTIC PHONOLOGICAL APPROACH **(With stress marker) (768X 1) XLNet embeddings**	**TRAIN ACCURACY (%)**	**TEST ACCURACY (%)**
Plosives	75.36	75.15
Fricatives	60.21	57.90
Glides	53.52	52.61
Nasals	67.53	66.83
Traps	65.72	65.46
Liquids	65.44	64.02
Trills	73.07	71.28

Table 2 provides a detailed view of the phonological attributes analysis, conducted using XLNet-transformer embeddings with dimensions of 768 x 1 per spoken word category for 10 subjects. The results demonstrate the accuracy of phoneme identification, revealing that fricatives achieved 57.90%, plosives reached 75.15%, glides scored 52.61%, nasals achieved 66.83%, traps attained 65.46%, liquids scored 64.02%, and trills exhibited an accuracy of 71.28%. Notably, plosives and trills outperformed other phonological attributes. This analysis underscores the significance of stress markers in linguistic understanding, as they play a crucial role in identifying spoken words through text transcripts. The EdgeCRNN model, introduced by Wei et al. (2021), which integrates feature enhancement through depth-wise separable convolution and residual connections, achieved an accuracy of 98.20%. Despite the demonstrated strength of Gammatone Frequency Cepstral Coefficients (GFCCs) in emotion detection, our approach outperformed the CNN-GFCC model proposed by Abdelmaksoud et al. (2021), which reported an accuracy of 93.09%. Similarly, our method surpassed the DenseNet-BiLSTM architecture—recommended for keyword spotting—with a reported accuracy of 94.88%. We also compared our system against other well-established models. For instance, in the domain of Urdu acoustic modeling, our approach outperformed the LSTM-based architecture proposed by Zia and Zahid (2019). Additionally, we achieved better results than the Deep CO-Training (DCT) algorithm reported by Cances and Pellegrini (2021). On the GSCD dataset, our method recorded a notable accuracy of 97.50%, exceeding that of attention-based encoder-decoder models such as the one by Higy and Bell (2018), which are known for their competitive performance. In the context of speech command recognition, our model demonstrated strong competitiveness when benchmarked against various state-of-the-art techniques. These included TripletLoss-res15 (Vygon & Mikhaylovskiy, 2021) at 98.38%, BC-ResNet-8 (Kim et al., 2021) at 98.70%, KWT-3 with self-attention (Berg et al., 2021) at 98.49%, RNN with neural attention (de Andrade et al., 2018) at 94.11%, MatchboxNet-3x2x64 (Majumdar & Ginsburg, 2020) at 97.63%, ConvMixer (Ng et al., 2022) at 98.21%, keyword spotting using Embedding + Head (Lin et al., 2020) at 97.70%, and Wav2KWS (Seo et al., 2021) at 98.52%. Notably, our method not only matches but often exceeds the performance of these models—particularly those based on transformer architectures—demonstrating its effectiveness and robustness for spoken command recognition across multiple benchmarks.

Table 4. Category-wise mathematical analysis of speech command

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

Our findings further suggest that higher-dimensional representations tend to yield more accurate results, contrasting the assumption that lower dimensionality is inherently beneficial. As shown in Table 3, the inclusion of phonemes with suprasegmental features—such as stress markers—led to a notable accuracy of 90.20%. In parallel, the visual modality, represented by mel spectrograms processed through EfficientNetV2 and our custom neural network, achieved an impressive accuracy of 99.78%. When we integrated the outputs of EfficientNetV2 with phoneme embeddings generated via XLNet, the combined system reached a peak accuracy of 99.80%. These results underscore the importance of incorporating stress markers within phonological features and highlight how high-dimensional embeddings enhance model performance. The synergistic fusion of phonetic information, suprasegmental cues, and visual speech representations through advanced deep learning models significantly boosts the accuracy of spoken word recognition.

Table 4 provides a detailed breakdown of precision, recall, and

F1-scores for each individual speech command category. Among these, precision serves as a crucial evaluation metric, measuring the proportion of correctly predicted positive instances out of all instances the model classified as positive, as defined in Equation (1). In essence, precision captures the model’s accuracy in identifying true positives while minimizing false alarms, offering valuable insight into how well the system distinguishes correct commands from incorrect ones.

$$precision = \frac{tp}{[tp+fp} \quad (1)$$

$$recall = \frac{tp}{[tp + fn]} \quad (2)$$

$$f1-Score = \frac{2 \times precision \times recall}{[precision+recall]} \quad (3)$$

Recall, also referred to as sensitivity or the true positive rate, serves as an essential performance metric. It is calculated as the ratio of true positives to the total number of actual positive cases, as shown in Eq. (2). This measure reflects the model’s effectiveness in identifying all relevant positive instances, with an emphasis on minimizing false negatives.

In this framework, true positives, true negatives, false negatives, and false positives are denoted by tp, tn, fn, and fp, respectively. To assess the model’s overall performance, we calculate the F1-score using the harmonic mean of precision and recall, as defined in Eq. (3). This metric is particularly valuable as it provides a balanced evaluation by incorporating both precision and recall into a single measure of accuracy.

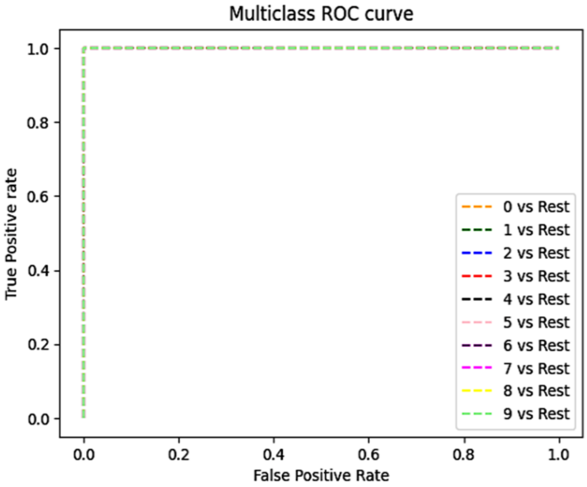

Fig 3: Multiclass evaluation ROC for classifying 10-word categories using EfficientNetV2

In our experimental setup, the 10-word categories demonstrated consistently high values across precision, recall, and F1-score metrics, indicating strong classification performance. As supported by the mathematical formulations, the statistical outcomes are summarized in Table 4. This table highlights that the majority of the classes were accurately classified, with only a few exhibiting slight deviations. Figures 3 presents the outcomes derived from our utilization of EfficientNetV2.

5. CONCLUSIONS

In conclusion, this research has effectively addressed the challenge of enhancing the precision of spoken word detection in ASR. We accomplished this by amalgamating multimodal data, specifically text transcripts and mel spectrograms, and employing the Speech2Text transformer to separate text from spectrograms, resulting in substantial improvements in spoken word identification accuracy. Through a comprehensive analysis of spectrograms and phonology, we gained valuable insights into speech articulation, contributing to a more robust understanding of linguistic data embedded in spoken speech. Our experiments, conducted on the Google Speech Command dataset version 2, utilized the ImageNet picture pool and second-generation EfficientNetV2 transformer for mel spectrogram image classification, affirming the efficacy of our proposed method. The integration of the G2P model further enhanced our comprehension of spoken speech by converting text transcripts into phonemes, allowing us to isolate specific phonological components. Ablation analysis was performed to assess the influence of spectrograms and phonological characteristics on the classification process, providing essential insights for system optimization. The late fusion technique, which combined phone embeddings and image embeddings, effectively extracted spoken words from rapid samples, underscoring the practicality of our approach. Notably, our experimental results exhibited substantial improvements in voice recognition accuracy compared to existing methods, establishing a new benchmark in spoken word recognition. By incorporating linguistic insights and leveraging diverse resources, our system achieved exceptional performance in ASR. Overall, this study represents a significant advancement in the field of automatic speech recognition, paving the way for future research in multimodal data analysis and promising more sophisticated and efficient speech processing systems.

6. REFERENCES

[1] Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy." *IEEE transactions on pattern analysis and machine intelligence* 41, no. 2 (2018): 423-443.

[2] Zhu, Hao, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. "Deep audio-visual learning: A survey." *International Journal of Automation and Computing* 18 (2021): 351-376.

[3] Sumby, William H., and Irwin Pollack. "Visual contribution to speech intelligibility in noise." *The journal of the acoustical society of america* 26, no. 2 (1954): 212-215.

[4] Lai, Kuo-Wei Kyle, and Hao-Jan Howard Chen. "An exploratory study on the accuracy of three speech recognition software programs for young Taiwanese EFL learners." *Interactive Learning Environments* (2022): 1-15.

[5] Nijhawan, Tanya, Girija Attigeri, and T. Ananthakrishna. "Stress detection using natural language processing and machine learning over social interactions." *Journal of Big Data* 9, no. 1 (2022): 1-24.

[6] Paula, Amauri J., Odair Pastor Ferreira, Antonio G. Souza Filho, Francisco Nepomuceno Filho, Carlos E. Andrade, and Andreia F. Faria. "Machine learning and natural language processing enable a data-oriented experimental design approach for producing biochar and hydrochar from biomass." *Chemistry of Materials* 34, no. 3 (2022): 979-990.

[7] BENSALAH, Rana Fadia, and Achouak BETTA. "The Impact of the Mother Tongue on the Phonetic Realization of Foreign Language Allophones. Algerian Arabic VS Received Pronunciation English." PhD diss., Université

Ibn Khaldoun-Tiaret-, 2022.

[8] Syafrizal, Syafrizal, Sri Wahyuni, and Tosi Rut Syamsun. "Pronunciation Errors of the Silent Consonants of Pariskian Junior High School Students." *Journal of English Language Teaching and English Linguistics* 7, no. 2 (2022): 155-165.

[9] Li, Bo, Ruoming Pang, Yu Zhang, Tara N. Sainath, Trevor Strohman, Parisa Haghani, Yun Zhu, Brian Farris, Neeraj Gaur, and Manasa Prasad. "Massively multilingual asr: A lifelong learning solution." In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6397-6401. IEEE, 2022.

[10] Li, Bo, Shuo-yiin Chang, Tara N. Sainath, Ruoming Pang, Yanzhang He, Trevor Strohman, and Yonghui Wu. "Towards fast and accurate streaming end-to-end ASR." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6069-6073. IEEE, 2020.

[11] Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in neural information processing systems* 33 (2020): 12449-12460.

[12] Chung, Yu-An, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training." In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 244-250. IEEE, 2021.

[13] Zhang, Yu, Daniel S. Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen et al. "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition." *IEEE Journal of Selected Topics in Signal Processing* 16, no. 6 (2022): 1519-1532.

[14] Graves, Alex. "Sequence transduction with recurrent neural networks." *arXiv preprint arXiv:1211.3711* (2012).

[15] Hu, Ke, Antoine Bruguier, Tara N. Sainath, Rohit Prabhavalkar, and Golan Pundak. "Phoneme-based contextualization for cross-lingual speech recognition in end-to-end models." *arXiv preprint arXiv:1906.09292* (2019).

[16] Reddy, Sravana, and James N. Stanford. "Toward completely automated vowel extraction: Introducing DARLA." *Linguistics Vanguard* 1, no. 1 (2015): 15-28.

[17] Jimerson, Robbie, and Emily Prud'Hommeaux. "ASR for documenting acutely under-resourced indigenous languages." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[18] Michaud, Alexis, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. "Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit." (2018).

[19] Omachi, Motoi, Yuya Fujita, Shinji Watanabe, and Matthew Wiesner. "End-to-end ASR to jointly predict transcriptions and linguistic annotations." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1861-1871. 2021.

[20] Li, Jinyu. "Recent advances in end-to-end automatic speech recognition." *APSIPA Transactions on Signal and Information Processing* 11, no. 1 (2022).

[21] Prabhavalkar, Rohit, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. "End-to-End Speech Recognition: A Survey." *arXiv preprint arXiv:2303.03329* (2023).

[22] Chiu, Chung-Cheng, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan et al. "State-of-the-art speech recognition with sequence-to-sequence models." In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4774-4778. IEEE, 2018.

[23] Chen, Xie, Yu Wu, Zhenghao Wang, Shujie Liu, and Jinyu Li. "Developing real-time streaming transformer transducer for speech recognition on large-scale dataset." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5904-5908. IEEE, 2021.

[24] Wang, Yuxuan, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang et al. "Tacotron: Towards end-to-end speech synthesis." *arXiv preprint arXiv:1703.10135* (2017).

[25] Shen, Jonathan, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779-4783. IEEE, 2018.

[26] Li, Naihan, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. "Neural speech synthesis with transformer network." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, pp. 6706-6713. 2019.

[27] Ren, Yi, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. "Fastspeech: Fast, robust and controllable text to speech." *Advances in neural information processing systems* 32 (2019).

[28] Ren, Yi, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. "Fastspeech 2: Fast and high-quality end-to-end text to speech." *arXiv preprint arXiv:2006.04558* (2020).

[29] Li, Naihan, Yanqing Liu, Yu Wu, Shujie Liu, Sheng Zhao, and Ming Liu. "Robutrans: A robust transformer-based text-to-speech model." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, pp. 8228-8235. 2020.

[30] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[31] Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training." (2018).

[32] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems* 27 (2014).

[33] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "Bart: Denoising sequence-to-sequence pre-training for natural language

generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019).

[34] Yi, Cheng, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. "Applying wav2vec2. 0 to speech recognition in various low-resource languages." *arXiv preprint arXiv:2012.12121* (2020).

[35] Gao, Heting, Junrui Ni, Yang Zhang, Kaizhi Qian, Shiyu Chang, and Mark Hasegawa-Johnson. "Zero-Shot Cross-Lingual Phonetic Recognition with External Language Embedding." In *Interspeech*, pp. 1304-1308. 2021.

[36] Chen, Tianlong, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. "The lottery ticket hypothesis for pre-trained bert networks." *Advances in neural information processing systems* 33 (2020): 15834-15846.

[37] Lai, Cheng-I. Jeff, Yang Zhang, Alexander H. Liu, Shiyu Chang, Yi-Lun Liao, Yung-Sung Chuang, Kaizhi Qian, Sameer Khurana, David Cox, and Jim Glass. "Parp: Prune, adjust and re-prune for self-supervised speech recognition." *Advances in Neural Information Processing Systems* 34 (2021): 21256-21272.

[38] Newell, Alejandro, and Jia Deng. "How useful is self-supervised pretraining for visual tasks?." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7345-7354. 2020.

[39] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[40] Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *IEEE Signal processing magazine* 29, no. 6 (2012): 82-97.

[41] Chen, Guoguo, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su et al. "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio." *arXiv preprint arXiv:2106.06909* (2021).

[42] Mohamed, Abdelrahman, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff et al. "Self-supervised speech representation learning: A review." *IEEE Journal of Selected Topics in Signal Processing* (2022).

[43] Chung, Yu-An, Wei-Ning Hsu, Hao Tang, and James Glass. "An unsupervised autoregressive model for speech representation learning." *arXiv preprint arXiv:1904.03240* (2019).

[44] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." *arXiv preprint arXiv:1807.03748* (2018).

[45] Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 3451-3460.

[46] Mehra, Sunakshi, Virender Ranga, and Ritu Agarwal. "Dhivehi Speech Recognition: A Multimodal Approach for Dhivehi Language in Resource-Constrained Settings." Circuits, Systems, and Signal Processing (2024): 1-21.

[47] Zhang, Qiuju, Hongtao Zhang, Keming Zhou, and Le Zhang. "Developing a Physiological Signal-Based, Mean Threshold and Decision-Level Fusion Algorithm (PMD) for Emotion Recognition." *Tsinghua Science and Technology* 28, no. 4 (2023): 673-685.

[48] Hazen, Timothy J. "Automatic alignment and error correction of human generated transcripts for long speech recordings." In *Ninth International Conference on Spoken Language Processing*. 2006.

[49] Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. "Xlnet: Generalized autoregressive pretraining for language understanding." *Advances in neural information processing systems* 32 (2019).

[50] Yenkimaleki, Mahmood, and Vincent J. van Heuven. "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 (2021): 102557.

[51] Duchi, John, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research* 12, no. 7 (2011).

[52] Warden, Pete. "Speech commands: A dataset for limited-vocabulary speech recognition." *arXiv preprint arXiv:1804.03209* (2018).

[53] Haque, Md Amaan, Abhishek Verma, John Sahaya Rani Alex, and Nithya Venkatesan. "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, 2020.

[54] Abdelmaksoud, Engy Ragaei, Arafa Hassen, Nabila Hassan, and Mohamed Hesham. "Convolutional Neural Network for Arabic Speech Recognition." *The Egyptian Journal of Language Engineering* 8, no. 1 (2021): 27-38.

[55] Wazir, Abdulaziz Saleh Mahfoudh Ba, and Joon Huang Chuah. "Spoken Arabic digits recognition using deep learning." In *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pp. 339-344. IEEE, 2019.

[56] Zia, Tehseen, and Usman Zahid. "Long short-term memory recurrent neural network architectures for Urdu acoustic modeling." *International Journal of Speech Technology* 22 (2019): 21-30.

[57] Lezhenin I, Bogach N, Pyshkin E (2019) Urban sound classification using long short-term memory neural network. In 2019 federated conference on computer science and information systems (FedCSIS), 57–60.

[58] Zeng, Mengjun, and Nanfeng Xiao. "Effective combination of DenseNet and BiLSTM for keyword spotting." *IEEE Access* 7 (2019): 10767-10775.

[59] De Andrade, Douglas Coimbra, Sabato Leo, Martin Loesener Da Silva Viana, and Christoph Bernkopf. "A neural attention model for speech command recognition." *arXiv preprint arXiv:1808.08929* (2018).

[60] Wei, Yungen, Zheng Gong, Shunzhi Yang, Kai Ye, and Yamin Wen. "EdgeCRNN: an edge-computing oriented model of acoustic feature enhancement for keyword spotting." *Journal of Ambient Intelligence and Humanized Computing* (2022): 1-11.

[61] Cances, Léo, and Thomas Pellegrini. "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, 2021.

[62] Higy, Bertrand, and Peter Bell. "Few-shot learning with attention-based sequence-to-sequence models." *arXiv preprint arXiv:1811.03519* (2018).

[63] Vygon, Roman, and Nikolay Mikhaylovskiy. "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, 2021.

[64] Kim, Byeonggeun, Simyung Chang, Jinkyu Lee, and Dooyong Sung. "Broadcasted residual learning for efficient keyword spotting." *arXiv preprint arXiv:2106.04140* (2021).

[65] Berg, Axel, Mark O'Connor, and Miguel Tairum Cruz. "Keyword transformer: A self-attention model for keyword spotting." *arXiv preprint arXiv:2104.00769* (2021).

[66] Majumdar, Somshubra, and Boris Ginsburg. "Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition." *arXiv preprint arXiv:2004.08531* (2020).

[67] Ng, Dianwen, Yunqi Chen, Biao Tian, Qiang Fu, and Eng Siong Chng. "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, 2022.

[68] Lin, James, Kevin Kilgour, Dominik Roblek, and Matthew Sharifi. "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, 2020.

[69] Seo, Deokjin, Heung-Seon Oh, and Yuchul Jung. "Wav2kws: Transfer learning from speech representations for keyword spotting." *IEEE Access* 9 (2021): 80682-80691.

[70] Mehra, Sunakshi, Virender Ranga, and Ritu Agarwal. "A deep learning approach to dysarthric utterance classification with BiLSTM-GRU, speech cue filtering, and log mel spectrograms." *The Journal of Supercomputing* (2024): 1-28.

[71] Mehra, Sunakshi, Virender Ranga, and Ritu Agarwal. "Improving speech command recognition through decision-level fusion of deep filtered speech cues." Signal, Image and Video Processing 18, no. 2 (2024): 1365-1373.

[72] Mehra, Sunakshi, Virender Ranga, Ritu Agarwal, and Seba Susan. "Speaker independent recognition of low-resourced multilingual Arabic spoken words through hybrid fusion." Multimedia Tools and Applications 83, no. 35 (2024): 82533-82561.

[73] Mehra, Sunakshi, 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.

[74] Mehra, Sunakshi, 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." Computational Intelligence 40, no. 6 (2024): e70012.