ISSN 2007-9737

Development of a Deep Ensemble Framework for Speech based Authentication System

Arpita Choudhury[1,*], Pinki Roy[1], Sivaji Bandyopadhyay[2]

[1] National Insitute of Technology Silchar,
Department of Computer Science and Engineering,
India

[2] Jadavpur University, West Bengal,
India

{arpitachoudhury_rs, pinki}@cse.nits.ac.in, sivaji.cse.ju@gmail.com

Abstract. Speech recognition has become ubiquitous on numerous platforms, such as AI-driven applications. Despite its extensive use, speech-based authentication remains less explored compared to traditional text-based methods. Every individual's voice is unique due to inherent anatomical differences. This uniqueness can be effectively employed in authentication applications by extracting and modeling these discriminative vocal features. Moreover, speech-based authentication is advantageous for hands-free operations and accessibility for individuals with disabilities. Integrating speech into a multi-factor authentication framework can significantly enhance security, making it more resilient against real-time fraudulent attempts. Nevertheless, achieving high speech recognition accuracy in the presence of intense background noise remains a major challenge. This work proposes a deep ensemble framework for speech recognition. The vVISWa and AudioMNIST datasets are used for experimentation. Both datasets are pre-processed to obtain audio samples of single-digit and digit sequence (6-digit) utterances. The audio samples are further augmented to mimic real-world conditions, such as extreme background noise and variation in speed and volume across speakers. The ensemble comprises three lightweight convolutional neural network (CNN) architectures: MobileNet, EfficientNetB0, and DenseNet121. The features extracted by these networks from mel-spectrogram representations of audio signals are integrated and fed into a squeeze-and-excitation block for recalibration. A fully connected softmax layer uses the recalibrated feature to perform classification. Under three-fold cross-validation, each model demonstrated significant performance gain. The average accuracy achieved by the ensemble model exceeds 95% for both the original(i.e. clean) and the augmented datasets. This makes the proposed system well-suited for deployment in secure, speech-driven authentication applications.

Keywords. Speech recognition, ensemble, deep learning, authentication.

1 Introduction

Deep learning-empowered AI applications with speech recognition technology are nowadays an integral part of smart human-machine interactive systems. Speech recognition enables machines to understand spoken words by converting the speech into text. This work advocates the use of speech recognition for hands-free authentication, such as unlocking applications with OTP(one time password) or verifying CAPTCHA challenges. Every individual's voice possesses unique characteristics that can be reliably modeled. Thus, further integrating speech recognition into multi-factor authentication systems with voice-based verification improves both security and user convenience. However, such applications remain limited in practice. Moreover, a major setback of real-time speech recognition is the background noise. Real-time noises are non-stationary due to their dynamic and time-varying nature. The CNN-based deep ensemble framework proposed in the present work

is inherently robust to noise, thereby eliminating the need for explicit noise removal.

The traditional acoustic features such as Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Perceptual Linear Predictive (PLP), and Relative Spectral Transform - Perceptual Linear Predictive (RASTA-PLP) [8, 7, 6] are compact, which causes loss of relevant spectral details, and they are more sensitive to noise. The present work instead utilizes mel-spectrograms that provide more detailed spectral information in the time-frequency domain. They also preserve temporal dynamics and capture nuanced characteristics of speech, which helps deep learning models learn sequential information effectively.

The deep ensemble framework (DEF) introduced in this paper recognizes speech accurately in diverse real-world acoustic conditions such as the presence of high intensity background noise. The ensemble comprises three deep learning classifiers EfficientNet-B0, MobileNet, and DenseNet-121 which are comparatively less resource-intensive than widely adopted models such as VGG, ResNet, and InceptionNet. Here, resource intensity indicates the computational power, memory, and storage requirements associated with model deployment and inference.

The organization of the rest of the paper is as follows. Section 2 outlines the recent trend in deep learning-based speech recognition, along with the associated challenges. The proposed methodology is elaborated in Section 3. The experimental results and their analysis are presented in Section 4. Finally, the limitations of the current work and future opportunities are discussed in the Conclusion section, i.e., Section 5.

2 Related Work

Speech recognition technology has been developing for several decades. The conventional methods for speech recognition lack efficiency in the presence of ambient noise [15], which affects the performance of real-time applications.

The introduction of deep learning has precipitated a significant paradigm shift in speech recognition research by enabling end-to-end learning. However, these models require high computational resources and a large amount of data. The present work studied the existing approaches to speech recognition and addressed the challenges related to noise-robust speech recognition.

Speech recognition is addressed in the literature through several paradigms, including phoneme recognition, isolated word recognition, continuous and spontaneous speech recognition.

Recognition of phoneme, the fundamental units of speech, is investigated by Palaz et al. [16] using normalized raw acoustic signals. Their CNN-based model with tanh activation function for feature learning outperforms handcrafted spectral features and ANN-based baseline architectures. The approach achieves a word error rate of 5.6% on the Wall Street Journal (WSJ) corpus. Feature invariance is further demonstrated through cross-domain evaluation on the WSJ and TIMIT datasets. Another phoneme recognition model employs attention-enabled transformer-based architecture with convolutional features [5]. The improved normalization technique recognizes utterances that are longer than the utterances present in the training data. The phoneme error rate achieved in this work is 17.6% using the TIMIT dataset.

P. Borde et al. [3] present an isolated digit recognition approach evaluated on the CUAVE and vVISWa datasets. The method employs LDA-normalized MFCC features with first- and second-order derivatives and a random forest classifier.The reported accuracies of 87.14% and 73.21% on CUAVE and vVISWa indicate modest performance compared to recent deep learning-based models. Sayed et al. [22] investigate the effectiveness of GFCC(Gammatone Frequency Cepstral Coefficients) features in comparison with MFCC for isolated English digit recognition under noisy conditions. The CUAVE dataset is augmented with Gaussian noise at multiple SNR levels, and performance is evaluated using LSTM(Long Short-Term Memory), SVM (Support Vector Machine), and DNN(Deep Neural Network) classifiers.

The SVM-based system achieves the highest accuracy among the evaluated models.

Nevertheless, GFCC features consistently underperform MFCC features, limiting their effectiveness for robust digit recognition.

CNN and RNN based deep neural networks has become more popular for speech recognition due to their capability of learning complex patterns from data. Revathi et al. [19] present a CNN-based framework for isolated word recognition evaluated on the TIMIT dataset. Multiple time–frequency representations, including spectrograms, mel-spectrograms, and gammatonegrams, are employed as features. The authors further extend the study to isolated Tamil digit recognition using data collected from hearing-impaired children.

The incorporation of phase spectrum compensation improves recognition performance, and the model is demonstrated on Raspberry Pi, highlighting its suitability for resource-constrained platforms. Swain et al. [24] propose a CNN and bidirectional attention LSTM-based framework for keyword recognition. Mel-spectrogram features extracted from denoised and augmented Google Speech Commands audio are used, with spectral gating applied for noise suppression. The reported accuracy of 92.94% indicates strong performance under controlled noise conditions. Sridhar et al. [23] compare multiple neural network architectures for digit recognition, including CNN, LSTM, Bi-LSTM (Bidirectional LSTM), GRU (Gated Recurrent Unit), and Conv-LSTM. Among the evaluated models, the Conv-LSTM architecture achieves the best performance, attaining 98.6% accuracy on the AudioMNIST dataset using MFCC features.

A deep convolutional network proposed for isolated English word recognition achieves 91.53% accuracy with MFCC features [21]. A residual network followed by a bi-directional GRU is found to be an effective approach for continuous speech recognition [20]. This approach replaces batch normalization with layer normalization.

The model achieves a character error rate of 4.7% and 3.61% respectively, when trained on LibriSpeech and LJ dataset.

A noise-robust end-to-end speech recognition model based on GRU achieves 98.0% accuracy in clean conditions and up to 14.1% gain in extremely noisy environment [18].

Recently, transformer architectures have also gained traction in the domain of speech recognition. Two transformer architectures, one with sequence-to-sequence (seq2seq) loss and the other with connectionist temporal classification(CTC) loss, are proposed for sentence recognition [1]. In this work, spectral magnitude is employed as the audio feature and the model is trained on the large-scale LRS2-BBC dataset introduced by the authors. The CTC model was found to be better than the seq2seq model in noisy environments in terms of word error rate (WER). A novel self-attention incorporated encoder-based framework called E-Branchformer is proposed in [12] for continuous English speech recognition using the LibriSpeech dataset. This model achieves a WER of 1.81%, which is better than the Conformer and Branchformer models developed in [17].

In [9], a noise-robust speech recognition framework is proposed using 1D convolution for feature extraction from the LRS2 and LRS3 datasets. Babble noise is superimposed on the audio samples to evaluate robustness within an attention-based encoder–decoder prediction framework . An extensive review of deep learning-based E2E models, including recurrent neural networks, transformers, and attention-based models for ASR is comprehended in [14].

The review of the literature shows considerable gaps in evaluating model performance under noisy conditions. Most research efforts seem to be focused on improving the accuracy of model performance for clean conditions while there is a comparative lack of comprehensive analysis of how such models actually function in complex real-world noisy settings. This is crucial for practical applications like automotive systems and mobile devices for health care and authentication. Despite the improved recognition rates with the advancements in deep learning models, most systems' accuracy is still insufficient for reliably functioning real-time applications. Moreover, high accuracy, is especially essential so that speech recognition technologies can smoothly be incorporated into everyday tasks where background noise is inevitable.

3 Methodology

3.1 Dataset

Two datasets vVISWa [4] and AudioMNIST [2] are used in this work. These datasets contain English digit speech data collected from multiple subjects. The vVISWa (Visual Vocabulary of Independent Standard Words) dataset is a multilingual dataset in which subjects utter standard words such as digits and city names across frontal, 45°, and side poses. The present work uses a subset of the vVISWa dataset, comprising 10 speakers (6 male and 4 female) who utter English digits. Each speaker in both datasets uttered the English digits from 0 through 9 in a repetitive manner. The AudioMNIST dataset comprises 30,000 audio recordings of spoken digits from 0 to 9, each approximately 1 second long at a 48 kHz sampling rate.

These recordings are collected from 60 participants (48 males and 12 females) aged between 22 and 61 with varied accents and origins. This dataset is recorded in a controled noise-minimal environment, organized by speaker directories. A metadata file is attached to the dataset detailing gender, age, and linguistic background for tasks like digit classification, speaker sex identification, or self-supervised learning benchmarks. The present work uses audio from 10 speakers (5 males and 5 females) from the AudioMNIST dataset. Each speaker repeatedly uttered each digit 50 times.

Mathematically, the data samples are represented as $\mathcal{S}_i = \{(S_a, S_v)_t\}_{t=1}^{T}$, where $t \in \mathbb{N}$ is a discrete-time index, S_a and S_v are the audio and video signal at time t.

After separating the audio and video streams, audio signals corresponding to single-digit utterances are first extracted. To simulate one-time password (OTP) scenarios, which typically consist of six digits, 20 predefined six-digit sequences are synthesized by concatenating individual digit recordings. During synthesis, sequences are constructed without repeating digits, and all merged digits originate from the same speaker to preserve speaker consistency. The processed audio signals are represented by $\mathcal{S}_a^{pr} = \{x_t\}_{t=1}^{T}$, where $t \in \mathbb{N}$ is a discrete-time index.

3.1.1 Data Augmentation and Pre-Processing

A range of data augmentation techniques is applied to the original and synthesized audio data. Clean audio signals are first superimposed with babble [13] and Gaussian white noise to simulate realistic environments, with signal-to-noise ratios set to -5 dB, 0 dB, 5 dB, and 10 dB. Babble noise represents a complex, non-stationary acoustic condition generated by overlapping speech from multiple speakers as shown in Figure 1. On the other hand, Gaussian white noise is stationary in nature i.e. it remains uniform across all frequencies. The audio volume is further adjusted by ±6 dB to account for variations in recording amplitude. Temporal augmentation is also performed by modifying the playback speed, increasing it up to 1.75 times and decreasing it to 0.5 times the original rate. As a result, each clean audio sample produces three augmented variants: noise-contaminated, volume-modulated, and speed-modified. This augmentation strategy contributes to a more diverse and robust dataset, thereby improving model generalization under real-world conditions. Let s_1,s_2,...s_n be the conversations of different individuals. Hence, babble noise can be modelled as $\psi = \sum_{i=1}^{n}(s_i)$.

3.1.2 Mel-Spectrogram Extraction

Mel-spectrograms are the visual representations of the audio signals in the time-frequency domain. Mel-spectrograms prevent information loss by preserving complete spectral information of audio signals, making them suitable for feature extraction by CNN-based architectures. In this work, mel-spectrograms are extracted from the original and noise-superimposed audio signals of the datasets.

The mel-frequency scale is adopted in mel-spectrograms, which represent the non-linear sound perception by humans appropriately. Mel-spectrograms are obtained by using equation (1) to compute the STFT(Short-Time Fourier Transform) of the signals first, followed by conversion to the mel scale using the equation (2):

$$S_a[u, v] = \sum_{i=0}^{n} S[v.H + n].win[n].e^{-j2\pi un/M}, \quad (1)$$

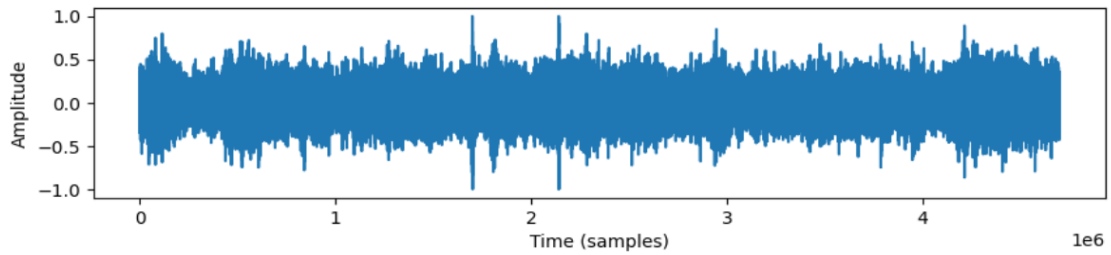

Fig. 1. Time-amplitude representation of babble noise

$$M_a(S_a[u,v]) = 2595log_{10}(1 + \frac{S_a[u,v]}{700}). \quad (2)$$

Finally, the mel-frequency components are converted to the logarithmic scale to plot the mel-spectrogram.

3.2 CNN Based Models

In this work, three CNN-based architectures EfficientNet-B0, DenseNet-121 and MobileNet are employed for feature extraction. These models are trained from scratch to learn noise-robust features from mel-spectrograms of original and noise-superimposed audio signals.

3.2.1 EfficientNet-B0

The conventional deep learning networks require random depth, width and resolution scaling to optimize the performance of the architecture. This manual tuning often results in minimal or zero improvement in efficiency. The authors in [25] devised compound scaling approach to mitigate this problem. EfficientNet-B0 is adopted as one of the classifiers in this work to perform.

Let, the EfficientNet-B0 contains i layers, then mathematically the output of EfficientNet-B0 can be defined as in eq. (3). V is the operation performed, N is the Convolutional layers and X is the input feature vector of height h_a, width w_a, and channels c_a:

$$Z_{Eff_{B0}} = V_i^{N_i}(X^{(h_a \times w_a \times c_a)}), \quad (3)$$

N depends on the coefficients depth d_{eff}, width w_{eff}, and resolution r_{eff} which are computed using (4). The EfficientNet B0 model uses $\alpha_{B0} = 1.2$, $\beta_{B0} = 1.1$, $\gamma_{B0} = 1.15$ and $\phi_{B0} = 1$ respectively:

$$d_{eff} = \alpha_{B0}^{\phi_{B0}}, w_{eff} = \beta_{B0}^{\phi_{B0}}, r_{eff} = \gamma_{B0}^{\phi_{B0}}. \quad (4)$$

The EfficientNet B0 architecture is presented in Figure 2. There are seven blocks in the architecture. Here, Conv represents the basic convolution, whereas MBConv refers to mobile inverted bottleneck convolution. MBConv reverses the conventional bottleneck structure of CNN by expanding the channel dimension. This step helps the model to learn complex features better. The digits 1 and 6 mentioned with MBConv are the expansion factors. A higher value of the expansion factor indicates a greater channel dimension. Also, a smaller kernel size (3×3) is ideal for learning local patterns, whereas larger kernels (5×5) capture the broad spatial context better.

3.2.2 DenseNet-121

The performance of the traditional CNN depends on the one-to-one connectivity between layers from input to output. DenseNet [11] on the other hand, maps each and every layer in a feed-forward way. It uses concatenation of feature maps instead of summation to exclude redundant features. Let x_a be the input image, $x_a \in \mathbb{R}^{(h_a \times w_a \times c_a)}$, where, h_a, w_a, c_a are the height, width, and channels. Let, f_0,f_1, ..., f_{n-1} be the feature maps from all preceding layers to layer n then the output concatenated features is given by

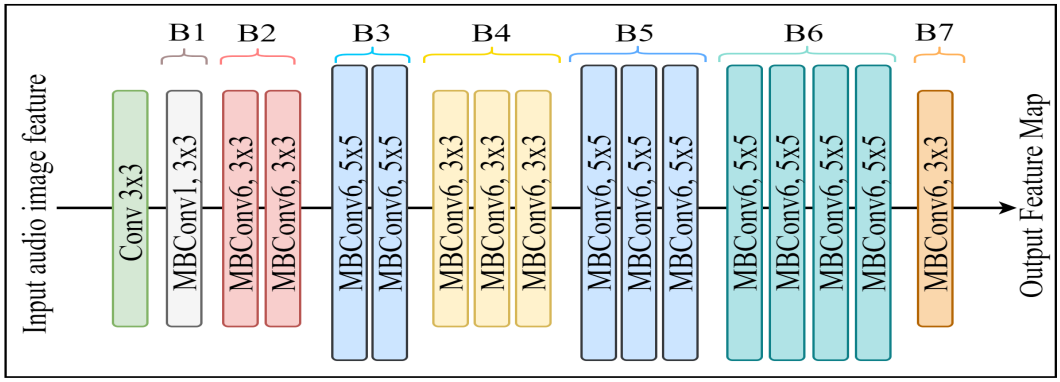

Fig. 2. EfficientNet B0 architecture, Conv refers to a convolution layer and MBConv refers to a mobile inverted bottleneck convolution. There are seven blocks from B1 to B7

f_n. Mathematically, the output can be illustrated using (5):

$$f_n = \mathcal{T}([f_0, f_1, ..., f_{n-1}]), \tag{5}$$

$\mathcal{T}$ is the composite operation of convolution, batch normalization, and ReLU. Further, the network is divided into dense blocks to perform viable concatenation operations as presented in Figure3. The dense blocks consist of a series of densely connected convolution layers. The four dense blocks present in DenseNet-121 architecture contain 6, 12, 24 and 16 layers respectively. There are transition layers between two dense blocks which use batch normalization, convolution and average pooling. Moreover, 1×1 convolutional layer is added in between dense blocks to reduce the number of feature maps efficiently. This layer is known as the bottleneck layer.

3.2.3 MobileNet

The MobileNet [10] is built on top of the factorization of standard Conv operation called depth-wise convolution (DW Conv) and 1×1 Conv called point-wise convolution (PW Conv). This 1×1 Conv is used to merge the outputs of the DW Conv. The DW Conv uses two different layers, one to filter the input and the other to combine the output. Let $\hat{V}$ be the DW Conv filters of size $U_v \times U_v \times C$ where U_v is the spatial dimension of the filter width and height, C is the number of filters associated with the operation. Hence, the DW Conv filter output ($\hat{O}_{m,n,c}$) is given by (6):

$$\hat{O}_{m,n,c} = \sum_{a,b} \hat{V}_{a,b,c} \mathcal{F}^{a}_{m+a-1,n+b-1,c}. \tag{6}$$

Each of the layers contains batch normalization and ReLU activation, which passes the merged output using 1×1 Conv. Finally, a fully connected layer with a softmax activation function is employed to classify the output class of the input feature vectors. Further, width and resolution multipliers α and ρ are added to reduce computational cost. This α is the scaler multiplied by the input and output channels. The α and ρ are set to 1 for this work. Figure 4 describes the MobileNet architecture.

3.2.4 Classification

The three backbone network independently processes the mel-spectrograms of size 224x224x3 and produce feature maps. These feature maps are subsequently passed through Global Average Pooling, which converts them into compact feature vectors of lengths 1024, 1280, and 1024. Each of these vectors is further reduced using a separate dense layer with ReLU activation, compressing them to 512 dimensions.

The reduced features from the three branches are then concatenated into a 1536-dimensional feature vector that integrates complementary information learned by different architectures.

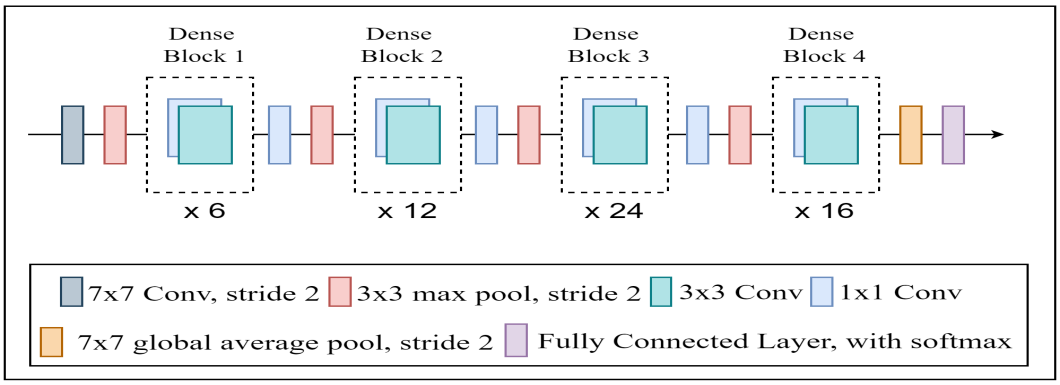

Fig. 3. Architecture of the DenseNet-121 model. The network consists of four dense blocks containing 6, 12, 24, and 16 densely connected convolutional layers. The blocks are separated by a transition layer composed of 1×1 convolution and pooling

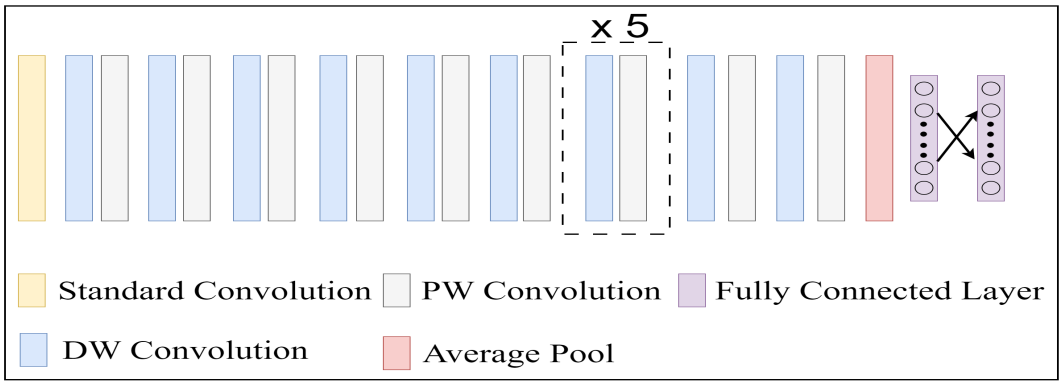

Fig. 4. Architecture of the MobileNet model based on depthwise separable convolutions. The network contains repeated stacks of depthwise (DW) and pointwise (PW) convolution layers

The fused vector is next passed to a Squeeze-and-Excitation (SE) block with a reduction ratio of 32 to apply channel-wise attention.

In this block, global contextual information is extracted to capture overall channel importance followed by a two-layer excitation network.

This network consists of a bottleneck layer with 48 units and ReLU activation, and an expansion layer with 1536 units and sigmoid activation. The resulting attention weights that highlight important channels are applied to the original fused features through element-wise multiplication.

The recalibrated features hence produced focus more on discriminative cues.The classification head operates on the recalibrated features representation. This head consists of a dense layer with 128 units and ReLU activation, followed by dropout regularization with a rate of 0.4 to reduce the chance of overfitting.

Finally, a softmax layer outputs the probability distribution over 50 audio classes.

Despite using multiple backbones and an attention mechanism, the overall model remains efficient, with a total of 16.2 million parameters.

This architecture shows that combining diverse feature extractors with adaptive attention can deliver strong audio classification performance without excessive computational cost. The

complete architecture of the proposed model, including all components, is presented in Figure 5.

4 Results and Discussion

4.1 Experimental Settings

The complete experiment is conducted on Kaggle, an open-source online platform with 16 GB of T4 GPU and 13 GB of RAM. The results and analysis of the proposed speech recognition system are described below.

The melspectrogram images extracted from the vVISWa and AudioMNIST dataset (both single digit and six digit data) are randomly split into three sets, training, validation and test set in the ratio 80:10:10. The number of samples in each set is presented in Table 1. Some of the sample melspectrograms obtained from the single digit vVISWa dataset are presented in Figure 6.The experiment is conducted three times for each dataset with independent train, test and validation sets to support statistical testing.

The batch size is set to 32 for all experiments, and the models are trained for 50 epochs. The number of trainable and non-trainable parameters in the proposed DEF is mentioned in the Table 2. Notably, these backbone architectures have some of the lowest parameter counts among popular pre-trained models such as VGG, InceptionNet, and ResNet.

4.2 Hyperparameter Tuning

The hyperparameters used for the mel-spectrogram extraction stage are presented in Table 3. The sampling rate of all the audio signals is 44100 Hz. Here, n_mfcc refers to the number of mel-spectrogram coefficients and n_mels refers to the number of filter banks considered for conversion to the mel-frequency domain.

The widely adopted hanning window function is employed while extracting mel-spectrogram to avoid spectral leakage. The number of audio samples between the beginning of consecutive frames is called hop length, which is set to 512 in this work. All the images in the dataset are further resized into $224 \times 224 \times 3$.

4.3 Performance Evaluation

4.3.1 Comparison with Traditional Models

The classification accuracies of a few traditional speech recognition classifiers from the literature are reported in Table 4. These classifiers include the support vector machine (SVM), Random Forest, long short-term memory (LSTM) network, and deep neural network (DNN), tested under varying noise conditions. These classifiers are trained using MFCC-13 and MFCC-39 feature representations obtained from both vVISWa and AudioMNIST single digit dataset.

The 13-dimensional Mel-frequency cepstral coefficients (MFCC-13) primarily capture static spectral characteristics of the audio signal, such as overall spectral shape and energy distribution. The 39-dimensional MFCC representation (MFCC-39) incorporates both static and temporal dynamics by including first-order (delta) and second-order (delta-delta) derivatives with the static coefficients. Experimental results indicate that the proposed DEF shows superior performance to the traditional classifiers across all conditions.

This performance gain is achieved at the cost of increased computational complexity. However, model performance and reliability are more important in security-critical applications.

4.3.2 Comparison with Backbone Models

The present work employed the three backbone models: EfficientNet-B0, DenseNet121, MobileNet for feature extraction. Additionally, the backbone models are individually trained after adding a classification head at the bottom of the network to compare their performance with that of the proposed model.

A detailed performance comparison of EfficientNet-B0, DenseNet121, MobileNet, and the proposed DEF model on vVISWa and AudioMNIST single digit datasets evaluated using mean accuracy, standard deviation, and 95% confidence interval (CI) is reported in Table 5. The CI and standard deviation are important measures

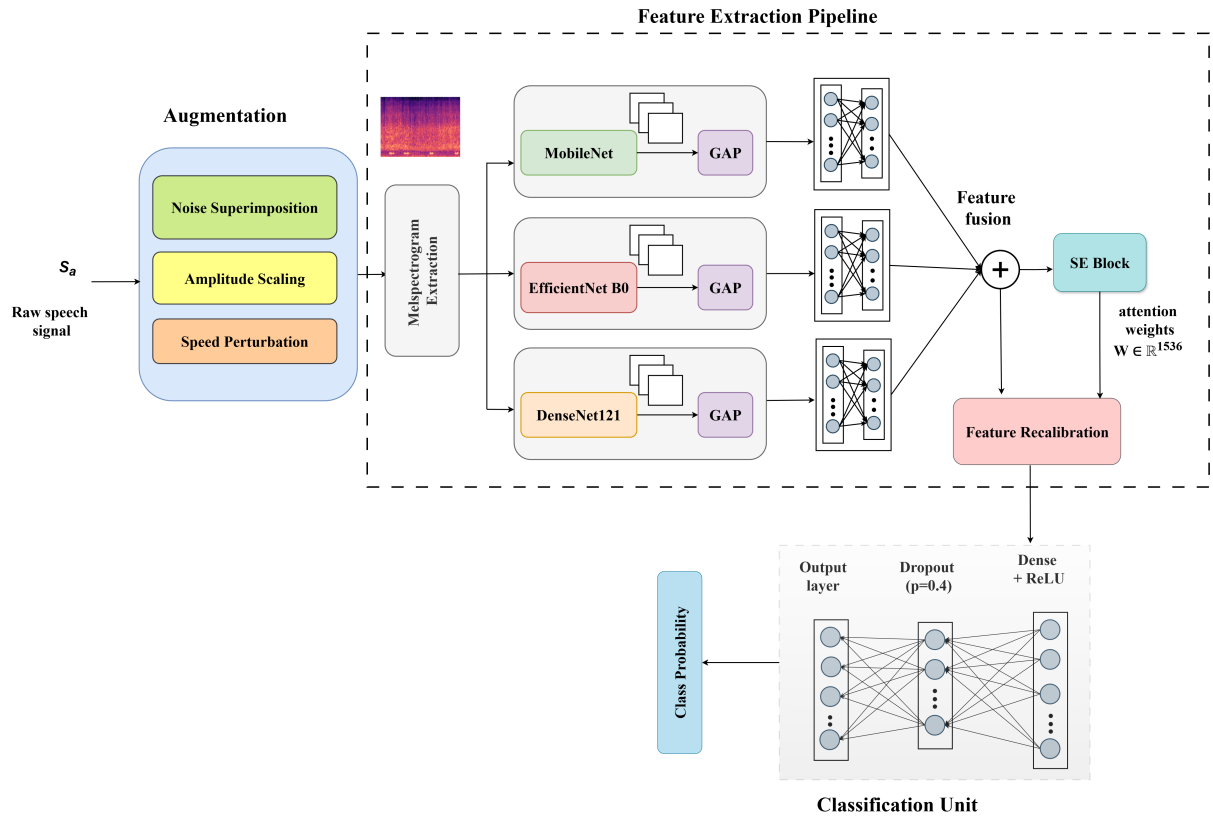

Fig. 5. Proposed deep ensemble framework for robust automatic speech recognition for authentication with three CNN based networks: EfficientNet B0, DenseNet121 and MobileNet as backbone

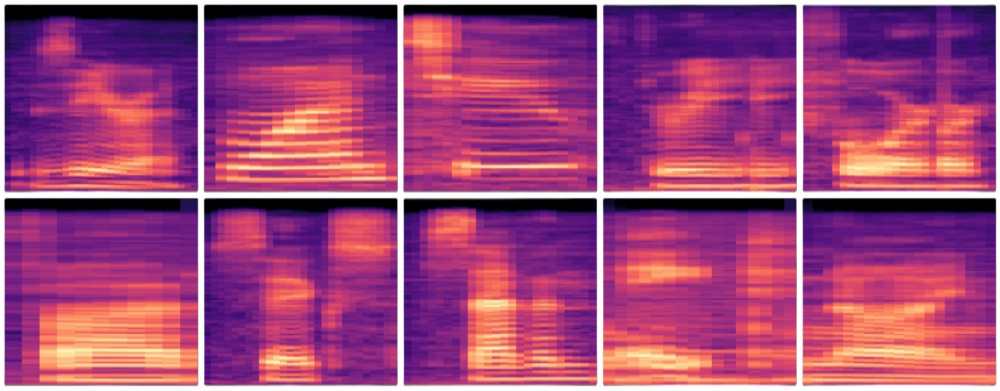

Fig. 6. Sample mel-spectrogram images extracted from vVISWa dataset for 0-9 English digit utterances

Table 1. Dataset distribution of training, testing, and validation samples for single-digit and six-digit sequence experiments using original and augmented versions of vVISWa and AudioMNIST datasets

	Dataset	Train	Test	Validation
Single digit	vVISWa_original	790	100	100
	vVISWa_augmented	3200	400	400
	AudioMNIST_original	4000	500	500
	AudioMNIST_augmented	32000	4000	4000
6 digit sequence	vVISWa_original	1600	200	200
	vVISWa_augmented	6400	800	800
	AudioMNIST_original	16000	2000	2000
	AudioMNIST_augmented	64000	8000	8000

Table 2. Number of parameters in the backbone and the proposed model

	Number of parameters		
Model	**Total**	**Trainable**	**Non-trainable**
EfficientNet B0	5,330,571	5,288,548	42,023
DenseNet121	7,302,474	7,218,826	83,648
MobileNet	3,493,834	3,471,946	21,888
Proposed DEF	18,105,737	17,916,178	189,559

for assessing a model's statistical robustness. The 95% confidence interval represents the range within which the model's actual performance is expected to lie with 95% confidence, based on repeated experiments or samples.

A narrower CI indicates more stable, reliable performance, while a wider CI indicates greater variability across runs. The equation for calculating 95% CI is presented in 7:

$$CI_{95\%} \approx \overline{acc} \pm 2.484 \times \frac{SD}{n}. \tag{7}$$

In eq. 7 $\overline{acc}$ refers to the mean accuracy achieved over the n independent runs, and SD represents the standard deviation. In the present work the value of n is set to 3, and the value 2.484 is the critical value from the Student's t-distribution. For the vVISWa-original dataset, the proposed model achieves the highest accuracy of 98.13%. Although DenseNet121 and MobileNet perform competitively, their wider CIs suggest greater variability compared to the proposed approach. Similarly, for the vVISWa-augmented dataset, the proposed model again produced the best performance with an accuracy of 98.64% and the smallest CI (± 0.0618). The relatively wider confidence intervals of EfficientNet-B0 and MobileNet imply less stable performance under augmentation.

On the AudioMNIST-original dataset, EfficientNet-B0 and DenseNet121 achieve strong baselines, yet the proposed model achieves the highest accuracy of 99.42%. with a substantially smaller standard deviation and a tight CI. Lastly, for the AudioMNIST-augmented dataset, although EfficientNet-B0 attains a competitive accuracy of 94.96%, the proposed model again dominates with 98.54% accuracy and the lowest variability.

Table 6 reports the performance of the backbone models and the proposed DEF model on the augmented vVISWa and AudioMNIST datasets for the task of classification of six-digit long digit sequences. The accuracy of all the backbone models on the augmented vVISWa dataset ranges from 95.44% to 96.39%. The proposed DEF model, on the other hand, exhibits the highest accuracy (97.49%), the lowest standard deviation (0.0058), with a very small confidence interval(± 0.0144). This shows that the proposed

Table 3. The hyperparameters and their values for feature extraction

Hyperparameter	Value
Samples per frame	2048
Hop Length	512
Windowing Function	Hann
n_mfcc	13
n_mels	128

Table 4. Comparative performance analysis of traditional speech recognition models using different audio feature representations on the vVISWa and AudioMNIST datasets. The table reports classification accuracy (%) obtained using MFCC13, MFCC39, and Mel-spectrogram features for both original and augmented versions of the datasets

Feature	Model	Dataset			
		vVISWa		AudioMNIST	
		Original	Augmented	Original	Augmented
MFCC13	SVM [22]	96.19	88.8	90.53	89.4
	Random Forest [3]	93.41	85.67	89.57	87.56
	LSTM [22]	85.66	82.28	86.35	84.6
	DNN [22]	90.54	88.54	90.53	89.14
MFCC39	SVM [22]	88.81	85.4	91.74	90.4
	Random Forest [3]	90.39	82.43	90.87	87.01
	LSTM [22]	88.07	77.91	84.1	96.78
	DNN [22]	90.35	88.61	92.1	91.87
Melspectrogram	Proposed Model	**98.13**	**98.64**	**99.42**	**99.54**

model not only improves the average performance, but it is also consistent.

The augmented AudioMNIST dataset for 6 digit long digit sequence showed the similar trend as the augmented vVISWa dataset. Although EfficientNet-B0 and DenseNet121 achieved high accuracy, their confidence intervals (± 0.1076 and ± 0.1120, respectively) remain broad. MobileNet exhibits larger CI (±0.2335) than other backbone models and the proposed DEF, which suggests that it is particularly sensitive to random data splits in sequence-level recognition. The proposed DEF model, with an accuracy of 97.33%, a lower SD (0.0332), and a narrower CI (± 0.0825), demonstrates model stability across diverse conditions.

The performance of the backbone models when trained separately for classification is hampered as their feature representations are limited by their distinct architectural biases, receptive fields, and optimization characteristics. Whereas the proposed DEF integrates these diverse backbones, by combining the complementing information acquired at various depths and sizes.

This integration mitigates model-specific mistakes, diminishes unstable activations, and improves feature robustness. The integrated framework exhibits enhanced generalization, less variance across iterations, and consistently superior performance relative to independent training of the backbone models.

Table 7 presents the inference time and floating-point operations (FLOPs) for the proposed DEF and for the backbone models as well, when they are individually trained.

Inference time refers to the wall-clock time taken by the trained model to process an input sample and produce an output during deployment.

Table 5. Performance comparison of the three backbone models and proposed models across the original and augmented vVISWa and AudioMNIST single digit datasets using classification accuracy, standard deviation (SD), and 95% confidence interval (CI)

Dataset	Model	Acc (%)	SD	CI
vVISWa_original	EfficientNet-B0	95.29	0.0564	± 0.1401
	DenseNet121	96.37	0.0407	± 0.1011
	MobileNet	96.00	0.0798	± 0.1983
	Proposed DEF	98.13	0.0352	± 0.0874
vVISWa_augmented	EfficientNet-B0	93.59	0.0876	± 0.2176
	DenseNet121	95.11	0.0451	± 0.1120
	MobileNet	94.37	0.0824	± 0.2047
	Proposed DEF	98.64	0.0273	± 0.0678
AudioMNIST_original	EfficientNet-B0	98.26	0.0552	± 0.1371
	DenseNet121	95.56	0.0840	± 0.2087
	MobileNet	95.56	0.0746	± 0.1852
	Proposed DEF	99.42	0.0174	± 0.0432
AudioMNIST_augmented	EfficientNet-B0	94.96	0.0634	$\pm$0.1575
	DenseNet121	94.40	0.0513	$\pm$0.1274
	MobileNet	96.19	0.0929	$\pm$0.2307
	Proposed DEF	98.54	0.0295	$\pm$0.0733

Table 6. Performance comparison of the three backbone models and proposed models across the augmented vVISWa and AudioMNIST for 6 digit data using classification accuracy, standard deviation (SD), and 95% confidence interval (CI)

Dataset	Model	Acc (%)	SD	CI
vVISWa	EfficientNet B0	96.58	0.0352	± 0.0875
	DenseNet121	95.44	0.0798	± 0.1983
	MobileNet	96.39	0.0188	± 0.0467
	Proposed DEF	97.49	0.0058	± 0.0144
AudioMNIST	EfficientNet B0	96.17	0.0433	± 0.1076
	DenseNet121	95.63	0.0451	± 0.112
	MobileNet	93.85	0.094	± 0.2335
	Proposed DEF	97.33	0.0332	± 0.0825

Table 7. The inference time and Flop count for the backbone models and the proposed DEF

Model	Inference time (ms)	FLOPs (G)
EfficientNet B0	2.2	0.39
DenseNet121	16	2.88
MobileNet	3.2	0.57
Proposed DEF	21.3	3.84

It is measured in milliseconds and is directly related to latency.

On the other hand, FLOPs represent the total number of arithmetic operations required to be performed during a single forward pass of the network. FLOPs is not hardware-specific; it estimates the computational complexity of a model.

Therefore, it is often used in literature to compare the theoretical efficiency of different models. Lower FLOPs and shorter inference times

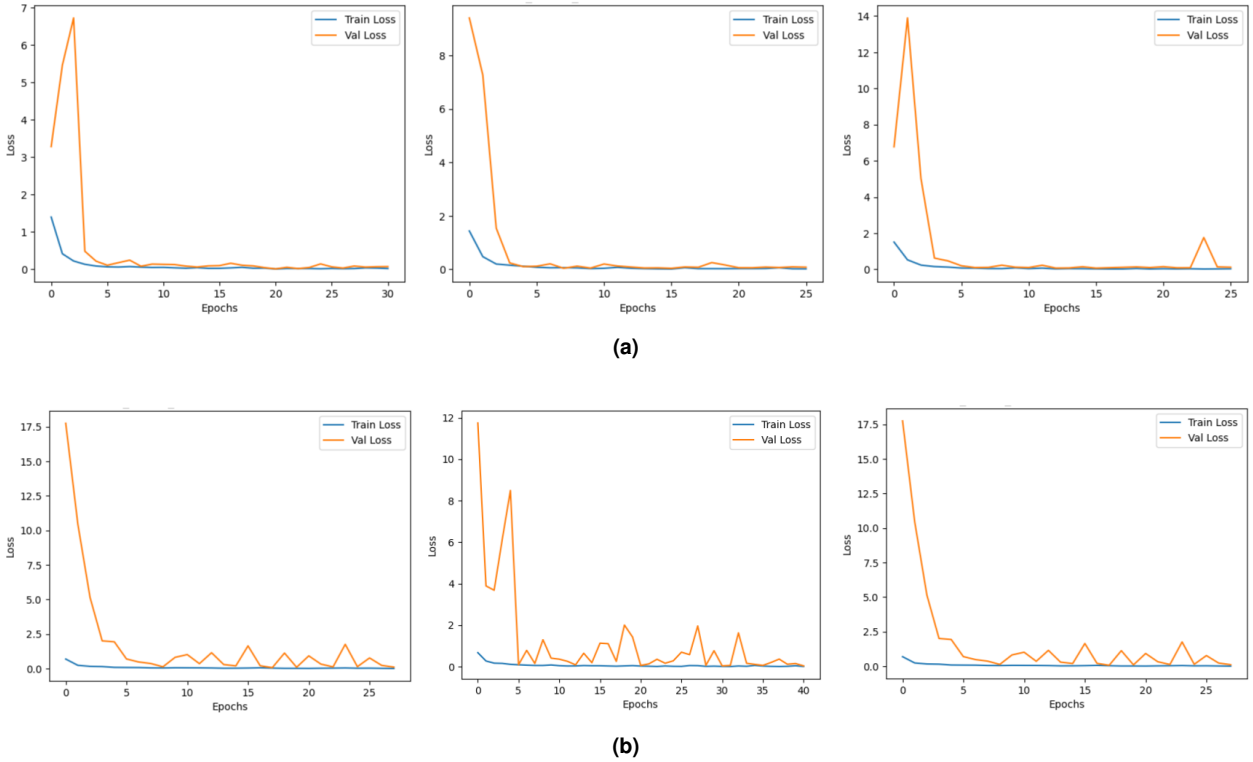

Fig. 7. The training vs. validation loss curves shown by the proposed DEF for (a) augmented 6-digit vVISWa dataset (b) augmented 6-digit AudioMNIST dataset, over three independent runs

generally indicate a more lightweight model with reduced computational and energy requirements. Such models are suitable for deployment on resource-constrained platforms. The proposed DEF exhibits higher FLOPs and inference time compared to its backbone models. This increase produces an additional computational overhead introduced by the proposed design.

However, this overhead with the consistent performance gain, indicates that the additional computations are effectively utilized to improve feature extraction and decision-making capability.

As the proposed DEF for an authentication system prioritizes accuracy and robustness, the higher FLOPs and inference time can be considered tolerable.

The loss curves of the model trained with the vVISWa and AudioMNIST datasets presented in Figure 7b also support the claim of consistency.

5 Conclusion and Future Work

Speech recognition has become an important aspect of effective human-machine interaction in our day-to-day lives. In the present work deep learning based robust speech recognition method suitable for hands-free OTP based authentication is proposed. The integration of the proposed speech recognition framework into a multifactor authentication system will enhance security and ensure user convenience simultaneously.

The proposed model is an ensemble of three CNN-based models: EfficientNet B0, DenseNet121 and MobileNet that act as the backbone of the architecture.

The model is trained with two datasets(vVISWa and AudioMNIST) that are augmented with noise, speed perturbation and volume scaling to ensure generalization. Mel-spectrogram

representations of the audio signals are fed to the model as input as mel-spectrograms present the detailed spectral information of the audio signals. The proposed ensemble achieves higher accuracies than traditional models and the individual backbone models for both original and augmented vVISWa and AudioMNIST dataset.

The statistical significance test performed based on standard deviation and confidence interval showed that the ensemble provides better optimization of speech recognition accuracy with consistency. Moreover, the low inference time and FLOPs count indicate that the model can be deployed in resource-constrained environments for real-time applications.

Despite the promising results of the proposed approach, there are some mentionable limitations. The present work considers speech recognition as a classification problem with predefines set of digit sequences. As the lack of temporal modelling affects the system's generalizability, introducing components such as LSTM and CTC (Connectionist Temporal Classification) loss will be helpful. A cross-modality experiment can be conducted using a recent dataset to elaborate on the robustness of the proposed approach. Extensive parameter tuning can be opted for to further optimize results in extremely noisy conditions.

Development of a complete multifactor authentication system in the future that integrates speech recognition with face recognition, iris recognition or voice recognition will firmly establish the claim.

References

1. **Afouras, T., Chung, J. S., Senior, A., Vinyals, O., Zisserman, A. (2018).** Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence, Vol. 44, No. 12, pp. 8717–8727. DOI: 10.1109/TPAMI.2018.2889052.

2. **Becker, S., Ackermann, M., Lapuschkin, S., Müller, K.-R., Samek, W. (2018).** Interpreting and explaining deep neural networks for classification of audio signals. arXiv preprint arXiv:1807.03418, Vol. 1. DOI: 10.48550/arXiv.1807.03418.

3. **Borde, P., Kulkarni, S., Gawali, B., Yannawar, P. (2022).** Recognition of isolated digit using random forest for audio-visual speech recognition. Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, Vol. 92, No. 1, pp. 103–110. DOI: 10.48550/arXiv.1807.03418.

4. **Borde, P., Manza, R., Gawali, B., Yannawar, P. (2004).** vviswa – a multilingual multi-pose audio visual database for robust human computer interaction. International Journal of Computer Applications, Vol. 137, No. 4, pp. 25–31. DOI: 10.5120/ijca2016908696.

5. **Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y. (2015).** Attention-based models for speech recognition. Advances in neural information processing systems, Vol. 28. DOI: 10.48550/arXiv.1506.07503.

6. **Dave, N. (2013).** Feature extraction methods lpc, plp and mfcc in speech recognition. International journal for advance research in engineering and technology, Vol. 1, No. 6, pp. 1–4.

7. **Debnath, S., Roy, P. (2020).** Appearance and shape-based hybrid visual feature extraction: toward audio–visual automatic speech recognition. signal image video process. DOI: 10.1007/s11760-020-01717-0.

8. **Desai, N., Dhameliya, K., Desai, V. (2013).** Feature extraction and classification techniques for speech recognition: A review. International Journal of Emerging Technology and Advanced Engineering, Vol. 3, No. 12, pp. 367–371.

9. **Hong, J., Kim, M., Choi, J., Ro, Y. M. (2023).** Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18783–18794. DOI: 10.1109/CVPR52729.2023.01801.

10. **Howard, A. G. (2017).** Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. DOI: 10.1109/CVPR.2017.243.

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

12. **Kim, K., Wu, F., Peng, Y., Pan, J., Sridhar, P., Han, K. J., Watanabe, S. (2023).** E-branchformer: Branchformer with enhanced merging for speech recognition. 2022 IEEE Spoken Language Technology Workshop (SLT), IEEE, pp. 84–91. DOI: 10.1109/SLT54892.2023.10022656.

13. **Krishnamurthy, N., Hansen, J. H. (2009).** Babble noise: modeling, analysis, and applications. IEEE transactions on audio, speech, and language processing, Vol. 17, No. 7, pp. 1394–1407. DOI: 10.1109/TASL.2009.2015084.

14. **Li, J., et al. (2022).** Recent advances in end-to-end automatic speech recognition. APSIPA Transactions on Signal and Information Processing, Vol. 11, No. 1. DOI: 10.1561/116.00000050.

15. **Meyer, J., Dentel, L., Meunier, F. (2013).** Speech recognition in natural background noise. PloS one, Vol. 8, No. 11, pp. e79279.

16. **Palaz, D., Doss, M. M., Collobert, R. (2015).** Convolutional neural networks-based continuous speech recognition using raw speech signal. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 4295–4299. DOI: 10.1109/ICASSP.2015.7178781.

17. **Peng, Y., Dalmia, S., Lane, I., Watanabe, S. (2022).** Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. International Conference on Machine Learning, PMLR, pp. 17627–17643. DOI: 10.48550/arXiv.2207.02971.

18. **Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., Pantic, M. (2018).** End-to-end audiovisual speech recognition. 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp. 6548–6552.

19. **Revathi, A., Sasikaladevi, N., Arunprasanth, D., Raju, N. (2023).** Raspberry pi-based robust speech command recognition for normal and hearing-impaired (hi). Multimedia Tools and Applications, pp. 1–25. DOI: 10.1007/s11042-023-17543-8.

20. **Reza, S., Ferreira, M. C., Machado, J. J., Tavares, J. M. R. (2023).** A customized residual neural network and bi-directional gated recurrent unit-based automatic speech recognition model. Expert Systems with Applications, Vol. 215, pp. 119293. DOI: 10.1016/j.eswa.2022.119293.

21. **Rudregowda, S., Patilkulkarni, S., Ravi, V., Gururaj, H., Krichen, M. (2024).** Audiovisual speech recognition based on a deep convolutional neural network. Data Science and Management, Vol. 7, No. 1, pp. 25–34. DOI: 10.1016/j.dsm.2023.10.002.

22. **Sayed, H. M., ElDeeb, H. E., Taie, S. A. (2023).** Bimodal variational autoencoder for audiovisual speech recognition. Machine Learning, Vol. 112, No. 4, pp. 1201–1226. DOI: 10.1007/s10994-021-06112-5.

23. **Sridhar, C., Kanhe, A. (2023).** Performance comparison of various neural networks for speech recognition. Journal of Physics: Conference Series, IOP Publishing, Vol. 2466, No. 1, pp. 012008. DOI: 10.1088/1742-6596/2466/1/012008.

24. **Swain, O. P., Hemanth, H., Saran, P., Kothandaraman, M., Ravi, L., Sailor, H., Rajesh, K. (2023).** Robust and efficient keyword spotting using a bidirectional attention lstm. International Journal of Speech Technology, Vol. 26, No. 4, pp. 919–931. DOI: 10.1007/s10772-023-10067-4.

25. **Tan, M., Le, Q. (2019).** Efficientnet: Rethinking model scaling for convolutional

neural networks. International conference on machine learning, PMLR, pp. 6105–6114. DOI: 10.48550/arXiv.1905.11946.

Article received on 12/08/2026; accepted on 10/09/2026.
Corresponding author is Arpita Choudhury.