RESEARCH ARTICLE

Feature Fusion for Performance Enhancement of Text Independent Speaker Identification

Zahra Shah[1,2,*], **Giljin Jang**[1] **and Adil Farooq**[3]

[1] School of Electronics Engineering, Kyungpook National University, Daegu 41566, Republic of Korea
[2] Sensify Inc., New York, NY 10016, United States
[3] Department of Electronic Engineering, Maynooth University, Maynooth, W23 A3HY, Ireland

Abstract

Speaker identification systems have gained significant attention due to their potential applications in security and personalized systems. This study evaluates the performance of various time- and frequency-domain physical features for text-independent speaker identification. Four key features—pitch (P), intensity (I), spectral flux (SF), and spectral slope (SS)—were examined along with their statistical variations (minimum, maximum, and average values). These features were fused with log power spectral features and trained using a Convolutional Neural Network (CNN). The goal was to identify the most effective feature combinations for improving speaker identification accuracy. The experimental results revealed that the proposed feature fusion method outperformed the baseline system by approximately 8%, achieving an accuracy of 87.18%.

Keywords: speaker identification, prosodic features, physical features, CNN, features fusion.

Academic Editor:
Prasun Chakrabarti

Submitted: 16 October 2024
Accepted: 09 December 2024
Published: 31 December 2024

Vol. 2, **No.** 1, 2025.
10.62762/TIS.2024.649374

***Corresponding author:**
✉ Zahra Shah
zahra.shah@sensifylife.com

1 Introduction

Speaker recognition, which utilizes speech signals to identify and verify a speaker's identity, is a crucial aspect of speech processing. It is divided into two primary applications: Speaker Identification and Speaker Verification [1]. Speaker identification is the process of determining the identity of an unknown speaker by comparing their speech signal with a database of known speakers. This involves a one-to-many comparison, where all enrolled voiceprints are evaluated in parallel to identify the most likely match [2]. In contrast, the speaker enrollment phase involves extracting and storing unique voice features in a database, while the identification phase performs parallel pattern matching to identify the speaker [1].

Speaker recognition systems can be classified into two categories: text-dependent and text-independent. In text-dependent systems, users are required to speak specific, pre-defined phrases. On the other hand, text-independent systems can recognize speakers regardless of the content they speak, offering greater flexibility but requiring more sophisticated algorithms. While text-dependent systems tend to offer higher accuracy in controlled environments, text-independent systems are more adaptable and suitable for real-world

Citation
Shah, Z., Jang, G., & Farooq, A. (2024). Feature Fusion for Performance Enhancement of Text Independent Speaker Identification. *ICCK Transactions on Intelligent Systematics*, 2(1), 27–37.

© 2024 ICCK (Institute of Central Computation and Knowledge)

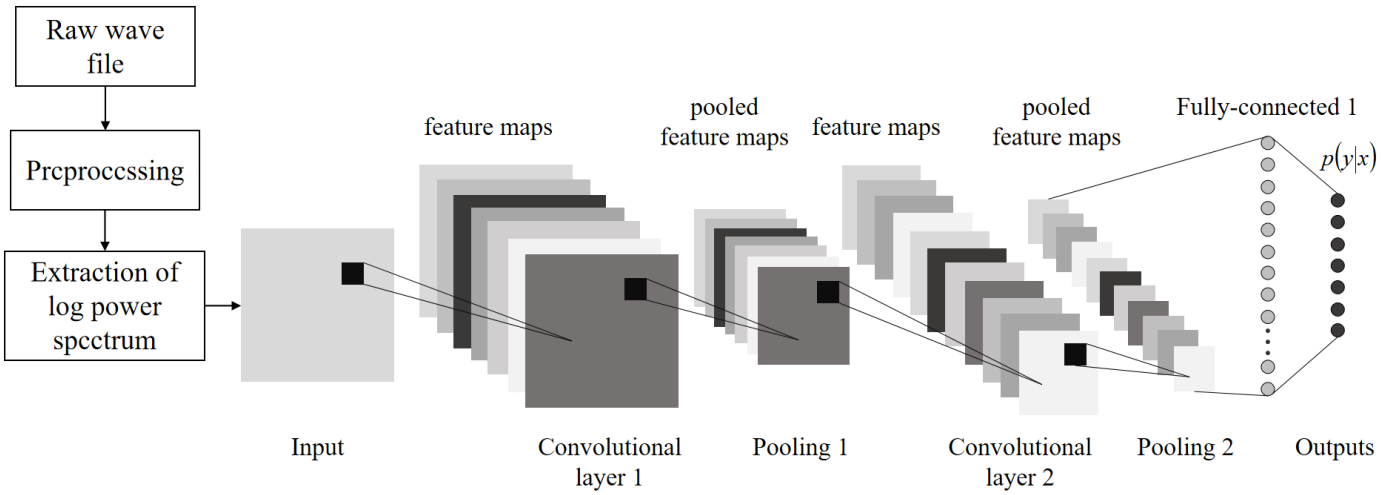


Figure 1. Baseline architecture.

applications.

Research in speaker recognition has evolved significantly, focusing on improving feature extraction techniques, developing advanced classification models, and enhancing speaker verification systems, especially in handling impostor models [4]. The growing need for robust, versatile, and efficient speaker recognition systems has driven substantial progress in this area. However, challenges remain, particularly in optimizing systems to handle large-scale datasets and diverse acoustic conditions [14, 15, 23, 27].

Although substantial research has been conducted in speaker recognition, there remains considerable potential for further advancements. There are many opportunities to optimize speaker recognition systems, including improving front-end processing, recognition algorithms, and speaker verification methods, as well as enhancing background models for unknown speakers [1]. Improvements can also be made in classification techniques or feature extraction methods to boost the system's performance [5]. Related advances in speaker diarization—which segments audio streams by speaker identity—offer complementary neural building blocks, such as the modular pipeline components introduced in Pyannote.audio, that can inform the design of more robust speaker identification front-ends [32]. These areas present exciting avenues for further improving the robustness and accuracy of speaker recognition systems.

In this study, we focus on enhancing the performance of text-independent speaker identification on a baseline architecture shown in Figure 1 by performing a comparative evaluation of various combinations of time- and frequency-domain features. Specifically, we explore Mel-frequency cepstral coefficient (MFCC), log-power spectral features, pitch, intensity, spectral slope, and spectral flux, in combination with convolutionally trained spectral features. Our objective is to evaluate the impact of these feature combinations on the accuracy of speaker identification systems. Experiments were conducted using the TIMIT dataset, where different feature combinations were tested to identify the most effective set for this task.

Additionally, we introduce a novel approach to text-independent speaker identification by employing feature fusion techniques to enhance the performance of CNNs. While earlier research has focused on individual features such as MFCCs, this work integrates acoustic, prosodic, and voice quality features to create a more comprehensive speaker representation. By combining these diverse features, we aim to improve speaker differentiation under varying conditions. The effectiveness of this feature fusion approach is demonstrated by its superior accuracy in speaker identification, which proves more robust to speech pattern variations compared to traditional methods [17, 18]. Further gains in discriminability can be achieved by applying margin-based training objectives to the embedding space, encouraging the model to produce more separable speaker representations [24].

2 Related Work

Feature extraction in speaker recognition involves both physical and perceptual features, which can be derived from the time domain or frequency domain. Some researchers have combined features

from both domains to improve system performance [6–8]. More recently, competitive speaker recognition systems have demonstrated that combining multiple subsystems and back-end scoring strategies—such as those submitted to the VoxCeleb Speaker Recognition Challenge—can yield further performance gains beyond what any single feature set achieves [25]. Classification techniques in speaker recognition commonly employ Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM), which are two of the most prevalent methods in the field. In addition, hybrid techniques combining HMM with Deep Neural Networks (DNN) have been explored, as well as other models such as neural networks, Vector Quantization (VQ), and combinations like HMM-NN and GMM-VQ. More recent approaches have utilized CNN and Recurrent Neural Networks (RNN) for classification [21], with end-to-end models learning directly from raw waveforms demonstrating competitive performance. Among the latest advancements, the i-vector technique has set new benchmarks in performance.

Bose et al. [7] have proposed a fusion of MFCC and Perceptual Linear Predictive Coefficients (PLPC) with an ensemble of classifiers, using MFCC-GMM as the baseline system. Their system models speakers using GMM and was evaluated on the NTIMIT dataset, achieving a maximum accuracy of 70.48%. Guo et al. [8] enhanced speaker identification by fusing subglottal resonance features with cepstral features. Their system uses a two-stage process: first, cepstral features are used to narrow down the candidate speakers, and then subglottal resonance features are employed with a multilayer perceptron to identify the target speaker. The study also investigated the impact of noise on performance, showing robustness under varying noise conditions.

In [15], the design and analysis of block-based transformations in MFCC computation were systematically evaluated for speaker recognition, demonstrating that careful feature engineering in the cepstral domain yields measurable gains in identification accuracy and robustness across varying acoustic conditions. The production of speech is divided into three stages: the excitation phase, which provides the power for speech production; the articulation phase, which involves the vocal tract shaping the sound; and the final phase involving the lips and nasal cavity. The energy from the lungs and the pressure produced during exhalation are key factors that influence sound volume and intensity; understanding these acoustic properties is foundational for deep learning-based speaker recognition systems that must model speaker-specific vocal characteristics derived from the speech production process [9].

The study in [16] explores the integration of MFCC with Power Normalized Cepstral Coefficients (PNCC), demonstrating enhanced robustness against background noise and significant improvement in Speaker Identification Accuracy (SIA). The work in [28] demonstrates that fusing heterogeneous and complementary feature streams— specifically prosodic features and short-term spectral features—through a two-stage classification architecture yields more discriminative speaker representations than any single feature type alone. This principle of complementary feature cooperation directly motivates our approach in speaker identification, where integrating spectral, prosodic, and voice quality features captures dimensions of speaker identity that no individual feature type can represent in isolation.

Feature extraction for speaker recognition tasks can be broadly classified into physical and perceptual categories [10]. Physical features represent the inherent characteristics of sound signals without considering how the sound was produced. These features can be extracted from various domains, such as time, frequency, or cepstral domains. For instance, dynamic features like zero-crossing rates, amplitude, intensity, loudness, and energy are all time-dependent characteristics that provide valuable information for speaker recognition. On the other hand, perceptual features, which are based on human auditory perception, help in distinguishing speakers. These features include pitch, fundamental frequency, jitter, shimmer, and tonality, and are crucial for speaker differentiation [11].

Perceptual features, inspired by the structure of human auditory and vocal organs, play a vital role in understanding speech signals. These features are influenced by the way speech is produced and processed, which varies depending on the species of the speaker, whether it be a human, animal, or bird. Prosodic features, often referred to as suprasegmental phonology, encompass qualities such as pitch, loudness, and rhythm, which allow humans to recognize speakers without consciously focusing on them.

Many studies have explored the use of MFCCs in isolation for speaker recognition, while others have

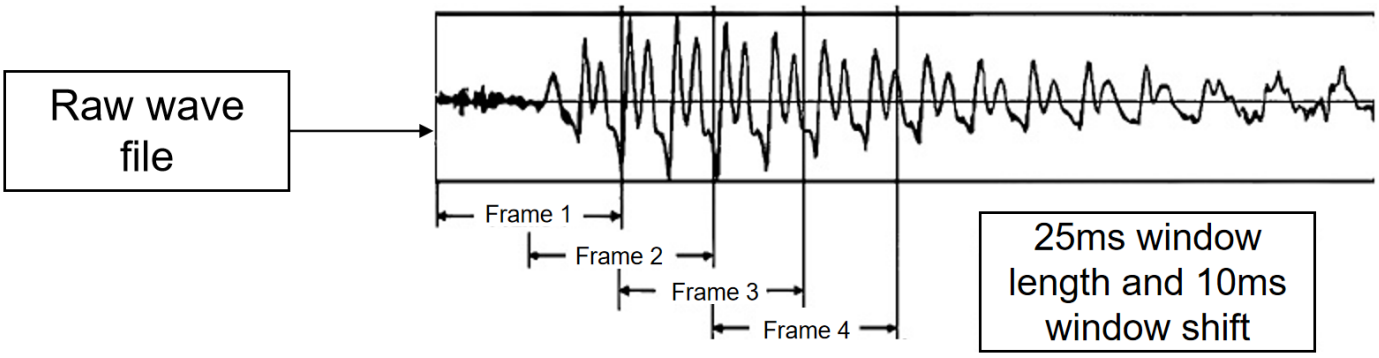


Figure 2. Preprocessing of speech signal (A).

moved toward learning compact speaker embeddings directly from data; for instance, x-vector systems trained with deep neural networks have demonstrated that discriminatively learned representations can substantially outperform hand-crafted feature pipelines [22]. Recent systems using deep neural networks and hybrid feature sets have achieved notable performance improvements by fusing spectral, prosodic, and voice quality features. In comparison, our method's unique feature fusion approach—integrating not only spectral but also prosodic and voice quality features—provides a more robust and accurate system for speaker identification. Moreover, our model's ability to perform well with diverse feature sets in challenging conditions highlights its superiority over traditional systems based solely on MFCCs or pitch alone.

3 Methodology

The methodology of this study involved several key steps to enhance the performance of the text-independent speaker identification system. The focus began with the front-end processing of speech signals, which was critical for extracting the most relevant features for recognition tasks. The first step in this process was the segmentation of the speech signal into overlapping frames. As shown in Figure 2, the speech signal was preprocessed by dividing it into overlapping frames, with a frame length of 25 ms and a hop length of 10 ms (60% overlap). This preprocessing method was crucial for extracting meaningful features from the raw speech signal [1, 2].

After segmentation, the extracted features were divided into time-domain and frequency-domain features, including pitch, intensity, spectral flux, and spectral slope, which had been proven to enhance speaker identification performance [3, 4]. These features were then combined with log-power spectral features, known for capturing key speaker characteristics such as voice quality and prosody [5, 6].

Once the features were extracted, a CNN model was employed to train the system using these features. The baseline system was based on a CNN architecture, which was trained on a combination of time- and frequency-domain features. The model was trained using a standard classification loss function, with accuracy as the primary performance metric. CNNs are particularly well-suited for learning hierarchical patterns in data and have been widely used in speaker recognition tasks. The performance of the system was evaluated by comparing the accuracy of speaker identification using various feature combinations.

In addition to feature extraction and model training, data augmentation techniques were implemented to improve the robustness of the model. This involved dividing each speaker's data into overlapping snippets of 90 frames, ensuring that all input data was of uniform length. This approach not only standardized the input but also increased the data for each speaker, enhancing the model's ability to generalize across different acoustic conditions [9]. The proposed system was tested on the TIMIT dataset, which provided a comprehensive collection of speech samples from a diverse set of speakers. The system's performance was evaluated in terms of classification accuracy, and the results were compared to baseline models to highlight the improvements made using the feature fusion and CNN-based approach.

After the raw speech signal was framed, the number of frames was determined based on the speech signal's length and the amount of overlap. Desired features were extracted from these frames for further analysis. Since the number of frames varied depending on the speaking speed and sentence length, each speech signal had a different number of frames. To train the model, it was necessary to ensure that each input had the same number of frames. The data augmentation

process described above is visually illustrated in Figure 3, which shows how overlapping snippets are extracted from each speaker's audio to standardize input lengths while introducing acoustic variability.

The features selected for this study included log-power spectral features (used as the baseline convolutional representation), and four physical features: pitch, intensity, spectral slope, and spectral flux. These were chosen for their ability to capture complementary speaker-specific characteristics across time and frequency domains. These features were chosen for their ability to capture key speaker-specific characteristics:

1. **Mel-frequency cepstral coefficients (MFCCs):** MFCCs are widely used for speaker recognition due to their ability to capture spectral properties that are robust across varying acoustic conditions [19].
2. **Prosodic features:** These include pitch and speech rate, which contribute to identifying speaker-specific intonation and rhythm. Such features have been shown to improve performance in noisy environments [20].
3. **Voice quality features:** Jitter and shimmer were selected as voice quality features. They provide additional speaker-specific characteristics, particularly useful in distinguishing speakers with similar speech patterns [19].

Combining these features enabled a more comprehensive representation of the speaker's vocal attributes.

After converting the speech signal into overlapping frames, several physical features were extracted. The four selected features—pitch, intensity, spectral slope, and spectral flux—were chosen for their relevance in speaker recognition tasks. Specifically:

1. **Pitch:** Harmonicity is critical to audio signal classification, and harmonic structure can be represented by multiple inharmonic peaks. The most common method for pitch detection is based on the autocorrelation function, where the highest value within a region of interest is used to detect pitch [12]. Given a discrete-time signal $x[n]$, the autocorrelation function $R_x(m)$ can be written as Equation 1:

$$R_x(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{+N} (x[n]x[n+m]) \quad (1)$$

2. **Intensity:** Intensity is primarily controlled by the force with which air from the lungs passes through the larynx. For our calculations, intensity was computed as energy per unit area, which is a common approach for measuring intensity in speech signals.
3. **Spectral Flux:** Spectral flux is the calculation of the 2-norm of two normalized power spectra. In our work, the log of the calculated spectral flux was used, as using the simple spectral flux did not yield the desired convergence in the system. The spectral flux was defined as Equation 2:

$$\text{spectralFlux} = \log\left(\text{l2-norm}\left(\text{PS}(\text{frame}[i]) - \text{PS}(\text{frame}[i+1])\right)\right) \quad (2)$$

4. **Spectral Slope:** The spectral slope is a measure of the spectral slant of a sound signal and is calculated using linear regression.

Pitch and intensity were extracted using the PRAAT Parselmouth library in Python [13]. Pretrained AlexNet was used for model training, with the 5th convolution layer fine-tuned using log power spectrum features. The physical features were then concatenated on the fully connected layer for the final speaker identification task.

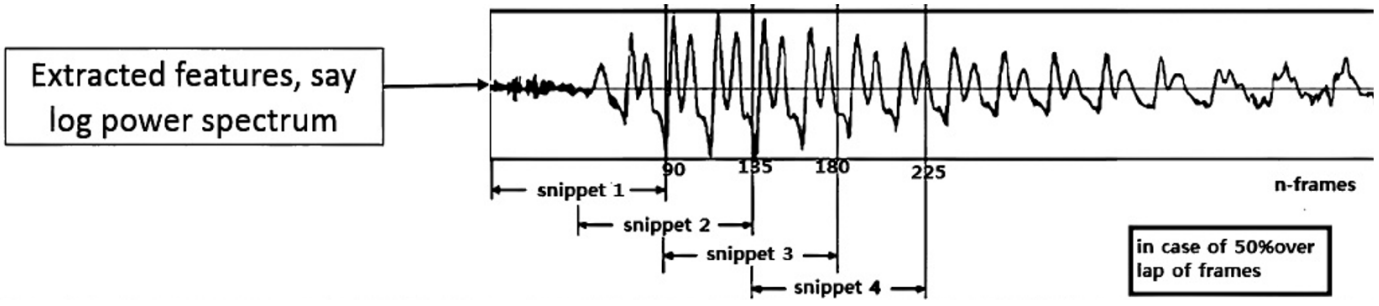


Figure 3. Preprocessing of speech signal (B).

Table 1. Dialect region distribution by gender.

Region	#Male	Male (%)	#Female	Female (%)	Total (%)
1	31	63%	18	37%	49 (8%)
2	71	70%	31	30%	102 (16%)
3	79	77%	23	23%	102 (16%)
4	69	69%	31	31%	100 (16%)
5	62	63%	36	37%	98 (16%)
6	30	65%	16	35%	46 (7%)
7	74	74%	26	26%	100 (16%)
8	22	67%	11	33%	33 (5%)
Total	**438**	**70%**	**192**	**30%**	**630 (100%)**

4 Dataset

The TIMIT dataset, utilized in this study, comprises 6,300 sentences spoken by 630 distinct speakers, including both male and female participants. Each speaker contributed 10 sentences, representing a diverse set of phonetic contexts. The speakers are distributed across eight distinct dialect regions of the United States, which reflect the geographical areas where they spent their childhood. These dialect regions are defined as follows:

- **dr1**: New England
- **dr2**: Northern
- **dr3**: North Midland
- **dr4**: South Midland
- **dr5**: Southern
- **dr6**: New York City
- **dr7**: Western (geographical boundaries poorly defined)
- **dr8**: Army Brat (frequent relocations during childhood)

These dialect regions were selected to capture a broad spectrum of linguistic diversity, which is crucial for evaluating the generalization capabilities of speech recognition systems. Notably, the dataset provides valuable insights into how speech characteristics vary across different regional accents and speaking patterns, thereby offering a robust framework for phonetic and dialectal research in speech recognition. The percentage inclusion of male and female speakers by the above dialect regions is shown in Table 1.

5 Experiments

The TIMIT dataset was used for our experiments, which contained 6300 sentences spoken by 630 different speakers, including both males and females. Each speaker provided 10 recordings. Ten recordings per speaker are a limited amount for training a speaker model. To address the issue of insufficient training data, we employed convolutional feature learning using MFCC and the power spectrum of the sound signal. To enrich the learned representations, phonetic structure was leveraged as an auxiliary signal during CNN training, following the approach of incorporating linguistic knowledge into convolutional speaker models [29]. We first extracted features from the raw audio file by performing a short-time speech signal analysis. After this analysis, the output consisted of several frames based on the length of the sound signal. Our experiments used a 25 ms frame length with a 10 ms hop length.

To mitigate the problem of having too little training data for each speaker, we sliced each feature vector into a specific number of frames. We evaluated the performance of our model with different numbers of frames and various amounts of frame overlap to achieve the highest performance. The numbers of frames used in the analysis were 15, 30, and 90 frames, with overlap percentages of 50% and 65%. The results obtained using log power spectrum features of the speech signal, with the mentioned overlap percentages, and employing fully connected layers as the recognition model, are shown in Table 2.

We extracted the log power spectrum of the sound signal using Python's 'python_speech_features' library. Pitch and intensity were extracted using the PRAAT Parselmouth library in Python [13].

Furthermore, to test the performance, we separated

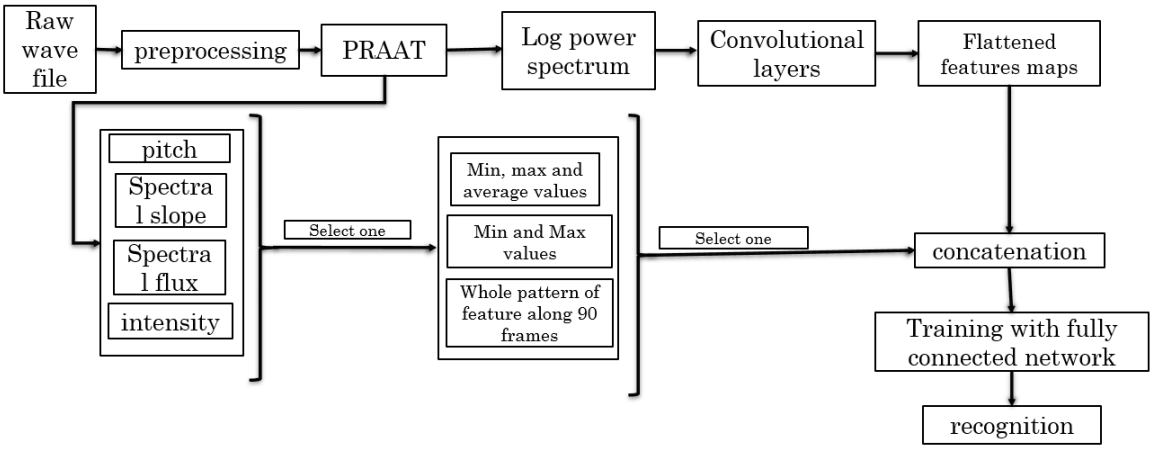


Figure 4. Model for concatenating individual features.

Table 2. Results with log power spectrum.

Number of frames	Frame overlap	Performance (%)
15	65	71.38
15	50	70.11
30	65	73.28
30	50	72.34
90	65	79.38
90	50	78.57

the training and test data for each speaker. Specifically, out of the 10 recordings of each speaker, we used 2 recordings as test data and the remaining 8 recordings for training the speaker model.

By examining the results with different frame overlaps and numbers of frames, it is evident that using 90 frames (equivalent to 2.25 seconds at 25ms per frame) yields the most reliable results for the speaker identification system. Additionally, a 65% overlap of frames provided the best performance for this system. When using only the log power spectrum, the highest performance achieved was 79.38%. In subsequent experiments, we used 90 frames as the standard for training, considering the effectiveness of this frame length.

Next, we discuss the concatenation of individual features, as shown in Figure 4, and their variations, as well as combinations of various features. While extracting other features such as pitch, intensity, spectral slope, and spectral flux, we observed that each speaker exhibited specific ranges of values for these features. Based on this observation, we decided to use these value ranges for further experimentation.

5.1 Results With Spectral Slope

We performed experiments with different variations of spectral slope concatenated to trained log-power spectral features, as shown in Table 3.

Table 3. Results with spectral slope.

Type of Feature	Frame overlap	Performance (%)
min, max, avg	50	83.65
min, max, avg	65	83.88
all values	50	83.11
all values	65	84.09
Min and max	50	83.57
Min and max	65	84.52

From the results, it can be concluded that each speaker has a very defined range of values for spectral slope, as the best results were achieved by concatenating the minimum and maximum values of the slope.

5.2 Results With Intensity

Similar to spectral slope, we performed experiments with different variations of intensity concatenated to trained log power spectral features, as shown in Table 4.

Table 4. Results with intensity.

Type of Feature	Frame overlap	Performance (%)
Min, avg, max	50	86.32
Min, avg, max	65	85.98
all values	50	84.38
all values	65	85.18
Min and max	50	83.81
Min and max	65	85.06

The best results were achieved by concatenating the

minimum, maximum, and average values of intensity, as indicated by the performance percentages for each variation of intensity values.

5.3 Results With Log Spectral Flux

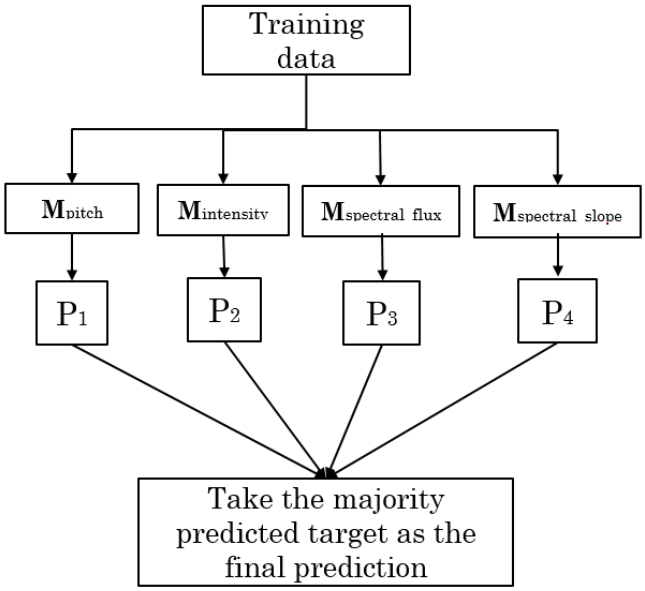


Figure 5. Process of majority vote.

Next, we performed experiments with different variations of Log Spectral Flux concatenated to trained log power spectral features, as shown in Table 5. Initially, we used simple spectral flux values, but the system failed to converge. After applying a logarithmic transformation to the spectral flux values, the system stabilized and achieved convergence, yielding a performance of 77%. After applying the log transformation to the spectral flux values, we obtained the following results.

Table 5. Results with log spectral flux.

Type of Feature	Frame overlap	Performance (%)
Min, max, avg	50	84.24
Min, max, avg	65	83.52
all values	50	82.98
all values	65	83.34
Min, max	50	83.11
Min, max	65	83.44

The best results were achieved by concatenating the minimum, maximum, and average values of log spectral flux, as indicated by the performance percentages.

5.4 Results With Pitch

We also performed experiments with different variations of Pitch concatenated to trained log power spectral features, as shown in Table 6.

Table 6. Results with pitch.

Type of Feature	Frame overlap	Performance (%)
Min, max, avg	50	81.18
Min, max, avg	65	82.23
all values	50	80.18
all values	65	80.98
min and max	50	81.38
Min and max	65	81.92

The best results were obtained by concatenating the minimum, maximum, and average values of pitch with a 65% frame overlap, achieving a performance of 82.23%, as shown by the performance percentages.

5.5 Results With Majority Voting

After performing experiments with individual features and their combinations, we conducted an experiment with majority voting to compare the results of our feature combinations with existing methods.

Majority voting is the process in which all models with individual features are used to predict the test data after training, as shown in Figure 5. The prediction given by the majority of the individual models is considered the actual prediction. The result of majority voting is shown in Table 7.

Table 7. Results with majority voting.

Type of Feature	Performance (%)
majority voting	84.54

5.6 Results With Different Combinations Of Features

After performing experiments with individual features, we also experimented with combinations of features, following Figure 6, to identify which combinations performed best for the task of text-independent speaker identification. The results are shown in Table 8.

5.7 Discussion

The findings of this study suggested that combining spectral slope and intensity features yielded favorable results in speaker recognition tasks. Specifically, the highest recognition accuracy (87.18%) was achieved by combining Intensity and Spectral Slope features with the CNN-trained log power

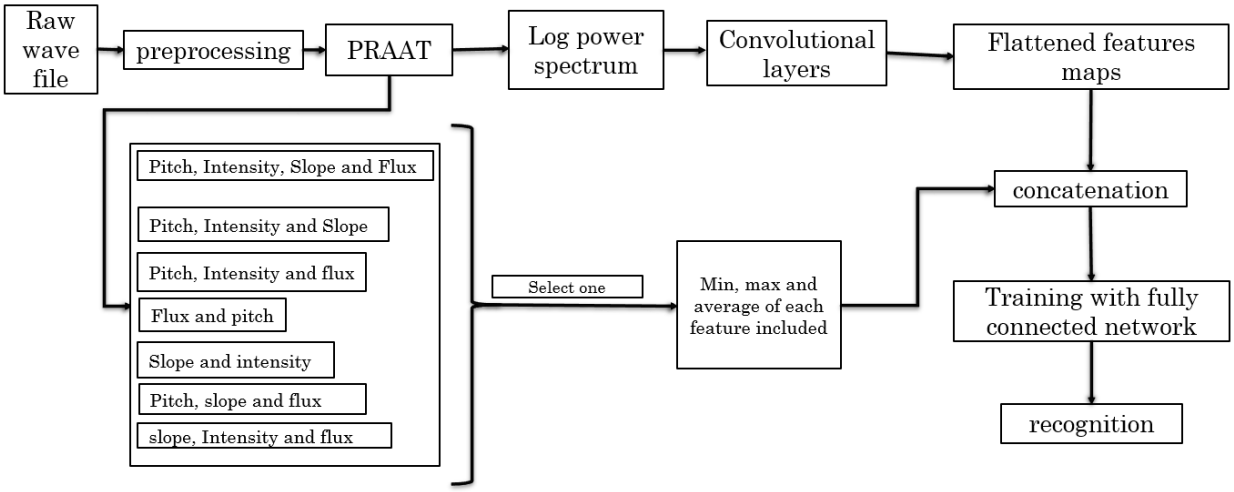


Figure 6. Model for combination of features.

Table 8. Results with combinations of features concatenated.

Type of Feature	Frame overlap	Performance (%)
P, I, SS, LSF	50	82.98
P, I, SS	50	85.38
P, I, SF	50	84.34
I, SS	50	87.18
P, LSF	50	83.44
P, SS, LSF	50	81.84
I, SS, LSF	50	84.85

spectral features. This outcome was attributed to the complementary nature of these feature sets. The CNN-trained log power spectral features captured detailed spectral information essential for speech sound characterization, while physical features such as intensity and spectral slope provided complementary speaker-specific cues that helped distinguish speakers based on vocal tract and energy characteristics. The synergistic use of both feature types allowed the model to generalize better across speaker variations, improving recognition accuracy. Combining spectral and prosodic features enhanced speaker discrimination by providing both phonetic (sound-based) and non-phonetic (tone-based) information, aiding differentiation in noisy or variable acoustic conditions.

The inclusion of voice quality features improved performance, especially in distinguishing speakers with similar-sounding voices. These features, sensitive to speaker-specific attributes like timbre and resonance, added differentiation not fully captured by MFCCs and prosodic features. Combining spectral, prosodic, and voice quality features enhanced the robustness of the speaker recognition system. Moreover, transformer-based architectures have shown strong capacity for modelling sequential and contextual dependencies in speech signals [30, 31]; their success in end-to-end speech processing tasks—including multi-speaker scenarios—suggests that similar attention mechanisms could be adapted to enhance the contextual modelling of speaker-specific features in identification systems.

6 Conclusion

This study demonstrated the scientific value of feature fusion techniques in enhancing the performance of text-independent speaker identification systems. The proposed method, which combined complementary features such as spectral slope and intensity, resulted in a significant improvement in system robustness and accuracy. By capturing both phonetic and prosodic speaker characteristics, the feature fusion approach provided a more comprehensive representation of the speaker's identity, crucial for real-world, noisy environments. The experimental results revealed that the fusion method outperformed the baseline system by 8%, achieving an accuracy of 87.18%. These findings highlight the potential of integrating spectral and prosodic features with advanced machine learning techniques to improve speaker recognition systems, especially in challenging conditions where traditional methods may struggle.

Although the feature fusion approach outperformed conventional methods, further investigations are needed to explore combining additional feature sets or advanced neural network architectures to enhance generalization. Future work could expand the dataset to include more speakers from diverse

acoustic environments to improve robustness. Additionally, exploring data augmentation techniques like generative models may boost performance in data-limited scenarios, and integrating advanced neural network architectures such as attention-based TDNN models [26] could further enhance speaker discrimination. These advancements would help speaker identification systems operate effectively in real-world conditions, supporting the ongoing development of speech recognition technologies.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

Zahra Shah is also affiliated with the Sensify Inc., New York, NY 10016, United States. The authors declare that this affiliation had no influence on the study design, data collection, analysis, interpretation, or the decision to publish, and that no other competing interests exist.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Sharma, R., Govind, D., Mishra, J., Dubey, A. K., Deepak, K. T., & Prasanna, S. R. M. (2024). Milestones in speaker recognition. *Artificial Intelligence Review*, 57(3), 58. [CrossRef]

[2] Mak, M. W., & Chien, J. T. (2020). *Machine learning for speaker recognition*. Cambridge University Press.

[3] Alrusaini, O., & Daqrouq, K. (2024). Text-independent speaker identification system using discrete wavelet transform with linear prediction coding. *Journal of Umm Al-Qura University for Engineering and Architecture, 15*(2), 112-119. [CrossRef]

[4] O'Shaughnessy, D. (2023). Review of Methods for Automatic Speaker Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*,vol. 32, pp. 1776-1789. [CrossRef]

[5] Shome, N., Sarkar, A., Ghosh, A. K., Laskar, R. H., & Kashyap, R. (2023). Speaker recognition through deep learning techniques: a comprehensive review and research challenges. Periodica Polytechnica Electrical Engineering and Computer Science, 67(3), 300-336. [CrossRef]

[6] Singh, M. K. (2024). A text independent speaker identification system using ANN, RNN, and CNN classification technique. *Multimedia Tools and Applications, 83*(16), 48105-48117. [CrossRef]

[7] Bose, S., Pal, A., Mukherjee, A., & Das, D. (2017). Robust speaker identification using fusion of features and classifiers. *International Journal of Machine Learning and Computing, 7*(5), 133. [CrossRef]

[8] Guo, J., Yang, R., Arsikere, H., & Alwan, A. (2017). Robust speaker identification via fusion of subglottal resonances and cepstral features. *the Journal of the Acoustical Society of America*, 141(4), EL420-EL426. [CrossRef]

[9] Bai, Z., & Zhang, X. L. (2021). Speaker recognition based on deep learning: An overview. *Neural Networks*, 140, 65-99. [CrossRef]

[10] Alías, F., Socoró, J. C., & Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, 6(5), 143. [CrossRef]

[11] Richard, G., Sundaram, S., & Narayanan, S. (2013). An overview on perceptually motivated audio indexing and classification. *Proceedings of the IEEE*, 101(9), 1939-1954. [CrossRef]

[12] Hui, L., Dai, B. Q., & Wei, L. (2006, May). A pitch detection algorithm based on AMDF and ACF. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* (Vol. 1, pp. I-I). IEEE. [CrossRef]

[13] Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71, 1-15. [CrossRef]

[14] Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language, 60*, 101027. [CrossRef]

[15] Sahidullah, M., & Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech communication, 54*(4), 543-565. [CrossRef]

[16] S. Al-Kaltakchi, M. T., Woo, W. L., Dlay, S., & Chambers, J. A. (2017). Evaluation of a speaker identification system with and without fusion using three databases in the presence of noise and handset effects. *EURASIP Journal on Advances in Signal Processing, 2017*(1), 80. [CrossRef]

[17] Xie, W., Nagrani, A., Chung, J. S., & Zisserman, A. (2019, May). Utterance-level aggregation for speaker recognition in the wild. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5791-5795). IEEE. [CrossRef]

[18] Ohi, A. Q., Mridha, M. F., Hamid, M. A., & Monowar, M. M. (2021). Deep speaker recognition: Process, progress, and challenges. *IEEE Access*, 9, 89619-89643. [CrossRef]

[19] Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*, 90, 250-271. pp. 250-271, 2017. [CrossRef]

[20] Koolagudi, S. G., Sreenivasa Rao, K., Reddy, R., Kumar, V. A., & Chakrabarti, S. (2012). Robust speaker recognition in noisy environments: Using dynamics of speaker-specific prosody. *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*, 183-204. [CrossRef]

[21] Ravanelli, M., & Bengio, Y. (2018, December). Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)* (pp. 1021-1028). IEEE. [CrossRef]

[22] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5329-5333). IEEE. [CrossRef]

[23] Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxceleB2: Deep speaker recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Vol. 2018, pp. 1086-1090). [CrossRef]

[24] Xiang, X., Wang, S., Huang, H., Qian, Y., & Yu, K. (2019, November). Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1652-1656). IEEE. [CrossRef]

[25] Zeinali, H., Wang, S., Silnova, A., Matejka, P., & Plchot, O. (2019). BUT System Description to VoxCeleb Speaker Recognition Challenge 2019. *arXiv preprint arXiv:1910.12592*.

[26] Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

[27] McLaren, M., Ferrer, L., Castan, D., & Lawson, A. (2016, September). The speakers in the wild (SITW) speaker recognition database. In *Interspeech* (pp. 818-822). [CrossRef]

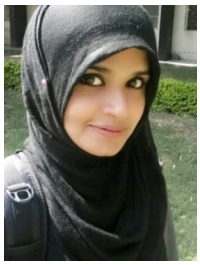
[28] Al-Qaderi, M., Lahamer, E., & Rad, A. (2021). A two-level speaker identification system via fusion of heterogeneous classifiers and complementary feature cooperation. *Sensors, 21*(15), 5097. [CrossRef]

[29] Zhou, T., Zhao, Y., Li, J., Gong, Y., & Wu, J. (2019, December). CNN with phonetic attention for text-independent speaker verification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 718-725). IEEE. [CrossRef]

[30] Liao, L., Afedzie Kwofie, F., Chen, Z., Han, G., Wang, Y., Lin, Y., & Hu, D. (2022). A bidirectional context embedding transformer for automatic speech recognition. *Information, 13*(2), 69. [CrossRef]

[31] Chang, X., Zhang, W., Qian, Y., Le Roux, J., & Watanabe, S. (2020, May). End-to-end multi-speaker speech recognition with transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6134-6138). IEEE. [CrossRef]

[32] Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., ... & Gill, M. P. (2020, May). Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7124-7128). IEEE. [CrossRef]



Zahra Shah received a B.S. degree in electrical engineering from the University of Engineering and Technology (UET), Taxila, in 2016. She received her MSc. degree from Kyungpook National University South Korea (2019). After her MSc., She worked as an AI researcher and developer in Ocean7 Corp NY. USA for 2 years. Then she worked as an AI researcher with a Pakistan-based startup for 1 year. Since January 2023 till date she has been working as a lead AI researcher and developer at Sensify Inc. NY, USA. (Email: zahra.shah@sensifylife.com)

Giljin Jang is a professor at Kyungpook National University, South Korea. He received his B.S. and M.S. degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea in 1997 and 1999, respectively. He also received his Ph.D. degree in the same department in February 2004. From 2004 to 2006, he was a research staff at the Samsung Advanced Institute of Technology, and, from 2006 to 2007, he worked as a research engineer at Softmax, Inc. in San Diego. From 2008 to 2009, he joined the Shiley Eye Center at University of California, San Diego as a postdoctoral scholar. From November 2009 to February 2014, he was an assistant professor at the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology (UNIST). (Email: gjang@knu.ac.kr)

Adil Farooq received an M.Sc degree in Electrical Engineering from the University of Engineering and Technology (UET), Taxila Pakistan, and a BS degree in Electronic Engineering from the International Islamic University Islamabad (IIUI), Pakistan. He has a decade-long experience working in both industry and academia. Currently, he is a doctoral researcher at Maynooth University Ireland working in the area of artificial intelligence and robotics. (Email: adil.farooq.2024@mumail.ie)