Int. j. inf. tecnol. (August 2024) 16(6):3493–3501
https://doi.org/10.1007/s41870-024-01877-z

ORIGINAL RESEARCH

A speaker identification-verification approach for noise-corrupted and improved speech using fusion features and a convolutional neural network

Rohun Nisa[1] · **Asifa Mehraj Baba**[1]

Received: 6 January 2024 / Accepted: 9 April 2024 / Published online: 19 May 2024

Abstract The degraded quality of the speech input signal has a negative impact on speaker recognition techniques. We address the issues of speaker recognition from noise-corrupted audio signals in the presence of four noise variants, including factory noise, car noise, street traffic noise, and voice babble noise, as well as noise-suppressed enhanced speech. The goal of this research is to create a speaker recognition algorithm that is resistant to a diverse spectrum of speech capture quality, background scenarios, and interferences. In this work, three distinct features, including Mel Frequency Cepstral Coefficients (MFCC), Normalized Pitch Frequency (NPF), and Normalized Phase Cepstral Coefficients (NPCC) are combined. The analysis that MFCC, NPF, and NPCC illustrate distinct features of speech underlies our observation. A Convolutional Neural Network (CNN) is used in our speaker recognition strategy to learn speaker-dependent attributes from fragments of Mel features, normalized pitch features, and phase cepstral features of clean speech, corrupted speech, and enhanced speech. The performance is measured using the ITU-T test signals and compared to previous algorithms at different Signal-to-Noise-Ratios of 0 dB, 5 dB, 10 dB, and 15 dB. For enhanced speech, all three features, MFCC, NPF, and NPCC, provided productive speaker identification and verification performance.

✉ Rohun Nisa
rohunnisa@islamicuniversity.edu.in

Asifa Mehraj Baba
asifa.baba@islamicuniversity.edu.in

[1] Department of Electronics and Communication Engineering, Islamic University of Science and Technology, Awantipora, Jammu & Kashmir, India 192122

Keywords Convolutional neural network · Feature extraction · Feature fusion · Speaker identification · Speaker verification · Speech enhancement

1 Introduction

Speaker recognition, a biometric method, utilizes speech features to authenticate a user's uniqueness through automated analysis of voice signals. Over recent decades, Automatic Speaker Recognition (ASR) systems have advanced significantly, finding applications in forensics, banking, and security. These systems comprise preprocessing, feature extraction, and speaker modeling components. Preprocessing involves refining input signals by eliminating non-speech elements and performing tasks like pre-emphasis and endpoint detection [1] [2]. Feature extraction, termed "front end preprocessing," transforms voice signals into numerical characteristics essential for training and testing speaker recognition systems. Speaker modeling constructs methods for speaker feature matching, crucial in the recognition stage for identification or verification purposes. Thus, speaker recognition systems serve vital roles across various domains, ensuring efficient and secure user authentication [3] (Fig. 1).

Speaker recognition systems often struggle in challenging acoustic environments due to factors like low audio SNR, diverse accents, and ambient noise, such as babble noise in crowded places. Conventional methods heavily rely on short-term spectral features like MFCC and Linear Prediction Cepstral Coefficients (LPCC), limiting their effectiveness in the presence of acoustic degradations. To address this, our research proposes a deep learning-based method called 1D-Frame Level-Feature Fusion-CNN. By combining MFCC with normalized pitch and phase features, this approach enhances recognition

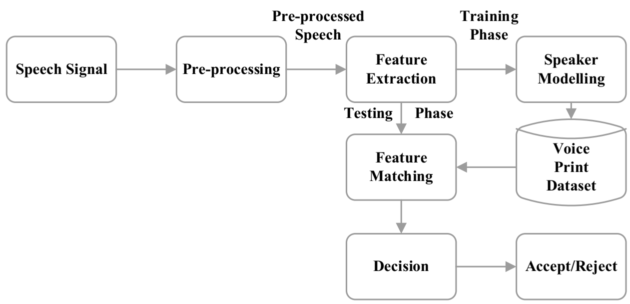

Fig. 1 A basic framework for an automated speaker recognition system

capabilities, even in scenarios with varying background noise strengths [4]. This research aligns with existing literature and offers promising advancements in speaker identification and verification techniques.

2 Overview of previous work

2.1 Earlier approach

Over the last decade, speaker recognition has undergone significant advancements, notably leveraging cepstral characteristics like MFCC [5]. Statistical and machine-learning methods such as Gaussian Mixture Model [6], Support Vector Machine (SVM) [7], and various score normalization techniques have been instrumental in speaker recognition systems. Recent improvements include the adoption of Gaussian Mixture Model-Universal Background Model (GMM-UBM) approaches [6], Support Vector (SV) techniques [8], and Factor Analysis-based engine voice (i-vector) architecture [9]. However, technical challenges persist in the domain, with environmental background noise and associated variations posing significant hurdles, particularly in scenarios with low signal-to-noise ratio.

The complex process of human speech involves various organs, yielding features indicative of pronunciation qualities in voice signals [10]. Speaker recognition algorithms integrate multiple speech characteristics to enhance accuracy [11]. Common feature extraction methods include LPCC, MFCC, Perceptual Linear Predictive Analysis, cepstrum differential coefficients, and RASTA filters [5, 12]. Spectrograms on the other hand offer a concise representation of acoustic features [13].

2.2 Deep learning approach

Recent advancements in speaker recognition, particularly with the adoption of deep learning, have significantly improved recognition rates and robustness [14]. MFCC, known for its resistance to noise and session variations, remains a cornerstone in this field [15]. Strategies for identifying similar MFCC feature vectors have been proposed [16], and CNN architectures have shown promise in enhancing accuracy [17]. Combining learned features with MFCC characteristics has yielded improved performance [18]. However, the computational demands of deep learning models remain a challenge [19], prompting the exploration of noise reduction techniques for robust speaker authentication.

Deep Neural Networks (DNNs) have demonstrated greater resilience to noise and acoustic reverberation compared to i-vectors, a machine learning approach incorporating GMM-UBM front-end with Probabilistic Linear Discriminant Analysis (PLDA) as the back-end classifier [20]. The benefits of voice-enhancing strategies with DNN embeddings in speaker recognition was investigated in [21]. El-Moneim et al. [22] focused on text-independent speaker recognition in noisy and reverberant environments, employing MFCCs, spectrum, and log-spectrum features analyzed by Long-Short Term Memory (LSTM)-Recurrent Neural Network (RNN) classifiers. Hourri et al. [23] proposed a novel method using CNN filters to extract speaker features, resulting in convVectors, which demonstrated enhanced performance under noise conditions.

A Two-level noise-robust PNN model (2LNR-PNN), addressing noise during preprocessing and feature extraction stages using spectral subtraction and GMM was introduced in [24], resulting in improved performance, reliability, and resilience in noisy and real-time scenarios. Hamidi et al. [25] utilized a Hidden Markov Model (HMM) based automatic speech recognition system to analyze cough signals, enabling the classification of coughs into sick or healthy category of speakers. AL-Shakarchy et al. [26] described a model designed to authenticate individuals based on their unique voice characteristics using deep learning techniques by leveraging the distinctive features present in their voices. Radha and Bansal [27] developed a child speaker identification system for non-native English speakers, evaluating fluency impact in text-dependent and text-independent tasks. Chelali [28] focused on audiovisual data fusion for robust speaker recognition in noisy environments, by extracting low-level features (LPC, MFCC for acoustic; ZM, HOG for visual) and fusing them to enhance modality efficiency.

3 Proposed methodology

Recognizing speakers in environments with minimal noise poses a challenge due to disruptions in crucial acoustical cues. This research aims to bolster system resilience and accurately identify desired speakers by simultaneously refining noise suppression techniques and speaker identification-verification procedures. This involves aligning learned features or enhanced speech signals with the necessary information for speaker identification-verification.

Speaker recognition involves identifying and confirming individuals based on their voice characteristics. This process facilitates tasks like personalized speech adaptation and speaker authentication for security purposes. Central to this process is feature extraction, which precisely characterizes speech signals amidst variations. By employing methods like Normalized Pitch information, MFCC, and Phase information, the feature extraction process translates acoustic input signals into patterns of acoustic feature vectors, providing a comprehensive depiction of speech signals. Deep Neural Networks are subsequently utilized to categorize speakers as either target or non-target based on these extracted features.

The proposed feature extraction process involves extracting cepstral features, Normalized Pitch Frequency, and phase information from the speech signal. Cepstral analysis, using Mel filter banks, decomposes speech signal frames into logarithmic spectral domain coefficients to model human ear effectiveness. Incorporating pitch frequency [29] with MFCCs [30] and phase information [31] aims to improve recognition outcomes. The classification process includes training and testing phases. During training, features from the enhanced speech signal train the Deep Neural Network model for each speaker. In testing, an unknown speaker's model is compared with learned features to decide on identification-verification. Speech signals from the ITU-T P-series recommendations directory [32] are used, with various real-world noise signals introduced before recognition.

3.1 Convolutional neural network processing

The presence of significant quantities of training data has driven primarily significant advancements in deep learning. However, such data is rarely available for particular tasks like speaker recognition, where significant amounts of information cannot be collected in real situations. As a result, in this work, we propose recognizing speakers using just a few training sets. To accomplish this, we employ a deep neural network with the Mel cepstral coefficients, normalized pitch spectrum, and phase cepstral coefficients as input, depicted in Fig. 2.

Our strategy for speaker recognition employs a CNN that is designed primarily to learn speaker dependent attributes from fragments of Mel features, normalized pitch features, and phase cepstral features of clean speech, corrupted speech, and enhanced speech. We developed a CNN-based feature level fusion method for combining and projecting speech attributes from the MFCC, Normalized Pitch, and Phase feature spaces into a d-dimensional joint feature space (explained in the later section). The value of d here is determined by the CNN architecture. The joint feature space is learned so that the joint feature representation encompasses highly discriminative speaker-dependent speech attributes, thereby enhancing speaker recognition accuracy. Before the feature extraction phase, we employ the speech enhancing method [33], to suppress the impact of noise on speech encountered in real-world scenarios. We deal with three instances including the clean speech, noise-corrupted speech and enhanced version of speech obtained from the method [33] for speaker identification-verification tasks. We will concentrate on text-independent speaker recognition throughout this work because it represents a more generalized form and has significant usage in a wide range of applications.

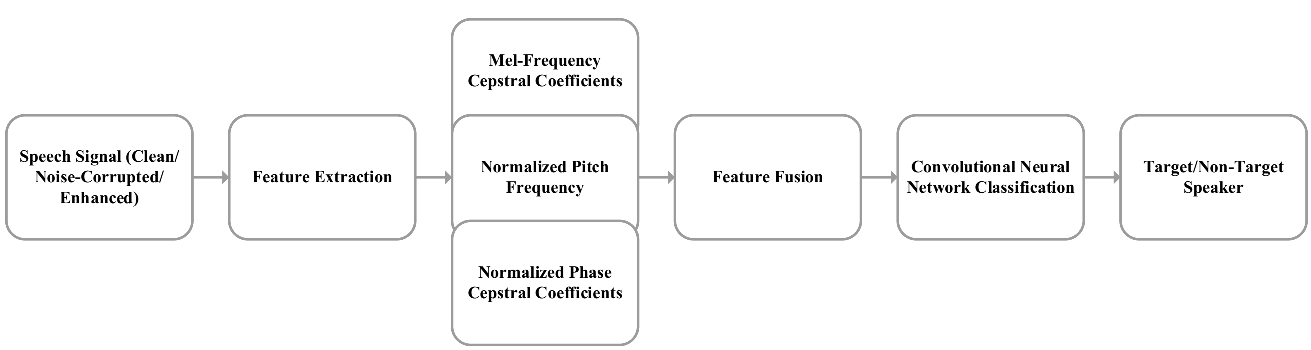

Fig. 2 Schematic overview of the proposed approach

4 Analysis of proposed method

The described procedures extract 40-dimensional Mel, normalized pitch, and normalized phase cepstral feature frames from speech frames. Each MFCC feature frame consists of 20 mel-cepstral coefficients (including the zeroth order coefficient), 20 first-order delta coefficients, 40 phase cepstral coefficients, and a normalized pitch. Cepstral Mean Variance Normalization (CMVN) is applied for feature normalization, enhancing generalizability in experiments. The number of speech frames obtained from a single voice file depends on the sampling frequency and voice length. For training the CNN with fixed-dimensionality input, 200 consecutive feature frames, termed "feature patches," are randomly sampled from each voice signal in every batch, resulting in feature patches of size 40×200. These MFCC, Normalized Pitch, and NPCC patches are stacked along the three-dimensional space to form a $40 \times 200 \times 3$ dimensional, three-channel feature patch named MFCC-NP-NPCC. Each channel represents MFCC, NP, and NPCC patches respectively. These features are integrated using 1D convolution filters in the CNN architecture, as illustrated in Fig. 3.

The CNN's objective is to transform each MFCC-NP-NPCC feature frame, with its 3-channel, 40-dimensional representation, into a 128-channel, 1-dimensional frame-level feature embedding. This 128-dimensional Joint Feature Space encapsulates speaker-dependent information linked to the input features. The arrangement of convolutional layers in a CNN significantly impacts its learning capability and effectiveness, as each layer learns distinct concepts from the data and refines information for deeper layers. ReLU non-linearity is applied to filter observations from each convolutional layer, mitigating the vanishing gradient problem commonly encountered with sigmoid activation functions. Additionally, max-pooling is employed to reduce the dimensionality of the network's learned space. Dropout layers are incorporated into the CNN during training to introduce regularization, offering the dual benefit of enhancing the CNN's resilience to input data variations while mitigating overfitting issues with the training data.

4.1 Speaker Identification Procedure

As shown in Fig. 3, during the testing phase, the input MFCC-NPF-NPCC feature strip $\mathcal{X}$, is divided into MFCC-NPF-NPCC patches, $\S_{\mathrm{i}}$, $\mathrm{i} \epsilon \{1, 2, \ldots, \mathcal{N}\}$, where $\mathcal{N}$, represents the number of patches. The CNN returns a series of classification scores, $\left\{\int_{\mathrm{i,j}}\right\}$, $\mathrm{j} \epsilon \{1, 2, \ldots, \mathcal{S}\}$, for each input MFCC-NPF-NPCC patch, $\S_{\mathrm{i}}$, pertaining to the $\mathcal{S}$ speakers. The classification score attributed to the jth speaker for the ith patch is represented by $\int_{\mathrm{i,j}}$. The combined classification scores, or $\{\mathcal{S}_{\mathrm{j}}\}$, for the complete speech signal are obtained by adding the results from each of the patches that were extracted from the speech signal, represented by Eq. 1, as:

$$\mathcal{S}_{\mathrm{j}} = \sum_{\mathrm{i}=1}^{\mathcal{N}} \int_{\mathrm{i,j}}, \forall_{\mathrm{j}} \tag{1}$$

The speaker j^* designated for the input speech signal is then chosen, given by Eq. 2, as:

$$\mathrm{j}^* = \underset{\mathrm{j}}{\operatorname{argmax}} \{\mathcal{S}_{\mathrm{j}}\} \tag{2}$$

4.2 Speaker verification procedure

For the verification of the intended speaker, Cosine Triplet Embedding Loss function is employed. In our scenario, we use the cosine similarity criterion, which offers superior learning dynamics than the Euclidean criterion and corresponds to the research in [32]. The cosine triplet embedding loss for training the model is represented by Eq. 3, as:

$$l\left(\mathcal{S}_{\hat{c}1}, \mathcal{S}_{\acute{n}1}, \mathcal{S}_{\hat{c}2}\right) = \sum_{\hat{c}1, \acute{n}1, \hat{c}2}^{\mathcal{N}} \mathrm{cosine}(\mathrm{f}(\hat{c}1, \acute{n}1)) - \mathrm{cosine}(\mathrm{f}(\hat{c}1, \hat{c}2)) \tag{3}$$

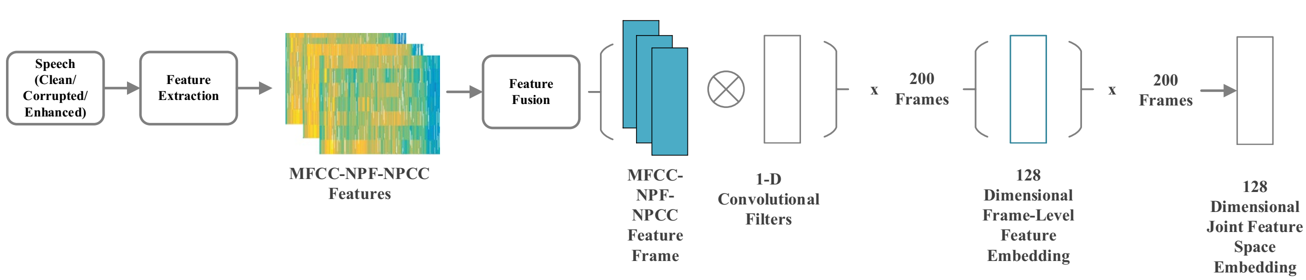

Fig. 3 A schematic of the proposed 1D-frame level-feature fusion-CNN architecture's feature fusion

Here, $l()$ represents the Cosine Triplet Embedding Loss function, $\hat{c}1$ corresponds to the clean speech utterance related to speaker 1, $\acute{n}1$ corresponds to the utterances of the noise-corrupted speech utterance related to speaker 1, and $\hat{c}2$ corresponds to the clean speech utterance related to speaker 2.

A whole essential criterion considered is that, despite the fact that the speech signal is changing constantly, the speaker-dependent vocal attributes are presumed to be quasi-stationary only over brief periods of time (15–35 ms) [34]. As a result, as mentioned in the feature extraction phase, we perform on short-term voice segments known as "voice frames". The MFCC, NPF, or NPCC features that correspond to that voice frame are referred to as the "Feature Frame". Therefore, a feature frame derived from a voice frame reflects just that voice frame's characteristics and has no correlation with its adjacent frames from the perspective of speaker recognition. Considering the above attribute limits, we specified the incorporation of 1D convolutional filters in conjunction with the feature frame in our Convolutional Neural Network for learning speaker-specific characteristic attributes of the concerned speech.

5 Experimental approach

5.1 Experimental setup

Our suggested approach was evaluated using the *ITU-T speech dataset*[1] [32] for clean speech signals. The collection from the ITU-T speech dataset contains 16 recorded sentences in every one of the 20 languages. Also, every set (or subset) contains half male speaker recordings and half female speaker recordings. The speech samples, initially at a 16 kHz sampling rate, were downsampled to 8 kHz to minimize the computational limitations of the system. The noise signals were chosen from the *NOISEX-92 dataset*[2] [35], including *factory noise*. Each of the resulting datasets was produced at one of three SNR levels: 0 dB, 5 dB, 10 dB, or 15 dB. For enhancing the speech signal, the method employed in [33] was incorporated as a preprocessing approach to deal with noise-corrupted speech samples from the speakers. Thus, three variants of speech signals, including clean speech samples, noise-corrupted speech samples, and enhanced speech samples, were provided as input to the proposed speaker recognition system. Because the text uttered by the speakers in the training and testing sets differ, the speaker recognition experimental studies are text-independent. Table 1 illustrates the performance evaluation of the proposed approach using state-of-the-art techniques.

5.2 Results and discussion

In this section, we evaluated the text-independent speaker recognition observation employing MFCC, NPF, NPCC information. Figures 4, 5, and Table 2 displays the results of the independent method for recognizing speakers in terms of Identification Accuracy (ID in %), Equal Error Rate (EER in %), False Acceptance Rate (FAR), and False Rejection Rate (FRR) for the factory noise scenario.

5.2.1 *Speaker identification results*

Figure 4 depicts the identification accuracy results of the proposed approach in comparison with the state-of-the-art techniques for factory noise, for noise-corrupted speech and enhanced speech.

Even when the SNR is 0 dB for factory noise, it is evident that the suggested method outperforms other baselines in different noise scenarios with accuracy of 40.5%, 41.6% for 5 dB, 42.8% for 10 dB, and 44.9% for 15 dB SNR variations. Furthermore, by incorporating an enhancement strategy, the proposed method improves identification performance even further with 93.8% identification accuracy at 0 dB, 94.7% at 5 dB, 95.4% at 10 dB, and 96.1% at 15 dB SNR levels, respectively. Before speaker identification, speech noise suppression is used, and a joint optimization is performed, which filters out some noise disruptions. The speaker-dependent speech improvement is implemented as well. With the exception of speaker-independent noise elimination, the incorporation of speaker knowledge not only recovers some of the noise-corrupted speech signals but also reveals speaker-specific characteristics that are important for speaker recognition.

5.2.2 *Speaker verification results*

Figure 5 depicts the speaker verification accuracy results of the proposed approach in comparison with the state-of-the-art techniques for factory noise, for noise-corrupted speech and enhanced speech.

The results of speaker verification are presented in the form of an Equal Error Rate (EER). The noise-corrupted and enhanced utterances are observed under four different SNR conditions. The proposed approach clearly benefits from

[1] https://www.itu.int/net/itu-t/sigdb/genaudio/Pseries.htm.

[2] https://svr-www.eng.cam.ac.uk/comp.speech/Section1/Data/noisex.html.

Table 1 Performance evaluation of the proposed approach using state-of-the-art techniques

Research study	Samples	Speech length	Features derived	Training/testing pairs	Method employed	Parameters evaluated	Recognition accuracy (in %)
Yutai (2009) [5]	Text dependent	Not mentioned	Dynamic MFCC	Not mentioned	Mel filters for pitch and MFCC GMM as classifier	Recognition rate	41.02 (10 dB) 73.28 (20 dB) 87.32 (30 dB)
Ali (2018) [18]	Text dependent	Urdu Dataset	Learned features + MFCC	10 speakers	DBN	Accuracy	92.6%
Siam (2019) [19]	Text dependent	ITU-T 4–12 s	MFCC	50 speakers	Spectral subtraction for noise + VQ for identification	Output SNR, recognition rate	36 (0 dB) 48 (5 dB) 68 (10 dB) 82 (15 dB) 94 (20 dB) 100 (25 dB)
Hourri (2021) [23]	Text independent	THUYG-20 SRE Corpus 30 s	MFCC + derivatives	371 speakers	RBM + UBM + CNN	EER DET	EER 1.05%
Juneja (2022) [24]	Text dependent	THUYG-20 SRE Corpus 4771 training utterances	MFCC, LPC, and statistical features	100M–100F/66M–87F	Spectral subtraction and GMM for noise + robust PNN model for identification	Accuracy, EER and FRR	Average accuracy 80% Maximum FRR 0.2
Proposed approach	Text independent	ITU-T 4-12 s	MFCC, NPF, PCC	10 speakers	CNN	EER, FAR, FRR, IDR	Maximum EER 2.25%, FAR 0.43205, FRR 0.11217, IDR 96.1% at 15 dB (enhanced)

speech noise suppression in all situations. The advantage is greatest at all the SNR levels, where the EER scores for noise-corrupted utterances are relatively high. At an input SNR of 0 dB, the EER score for noise-corrupted speech estimates to 16.11%, 14.21% for 5 dB, 10.39% for 10 dB, and 7.18% for 15 dB SNR levels. Using the proposed method for enhanced speech utterances, this score is reduced to 7.13% for 0 dB, 5.47% for 5 dB, 3.98% for 10 dB, and 2.25% for 15 dB SNR values, respectively. In comparison to prior speaker verification systems, the proposed method outperforms them in all conditions. As a result, using MFCC in conjunction with NPF and PCC enhances speaker verification performance uniformly over existing methods at all input SNR levels.

Table 2 displays the results of the performance assessment for noise-corrupted speech and enhanced speech under the influence of factory noise in terms of False Acceptance Rate (FAR) and False Rejection Rate (FRR). For factory noise, the presented approach successfully accomplished promising outcomes of 0.35942 FAR and 0.11096 FRR at

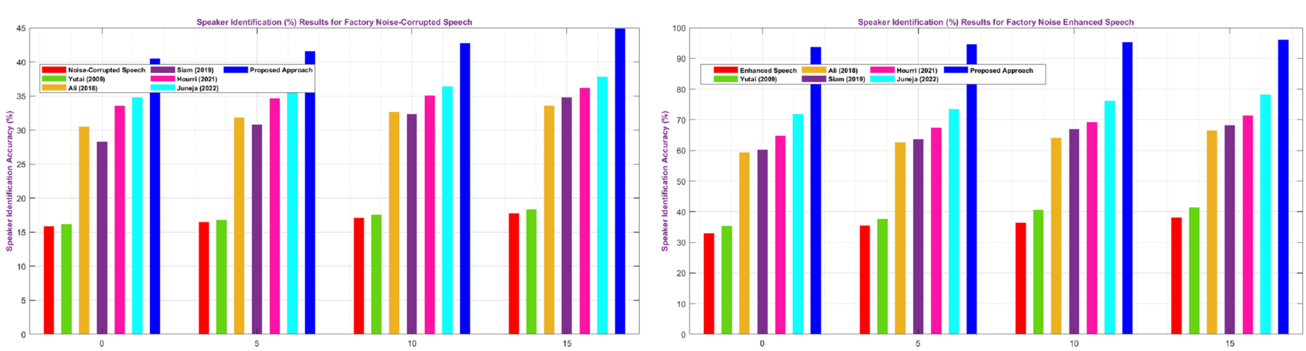

Fig. 4 Speaker Identification results (in %) for noise-corrupted speech and enhanced speech (factory noise)

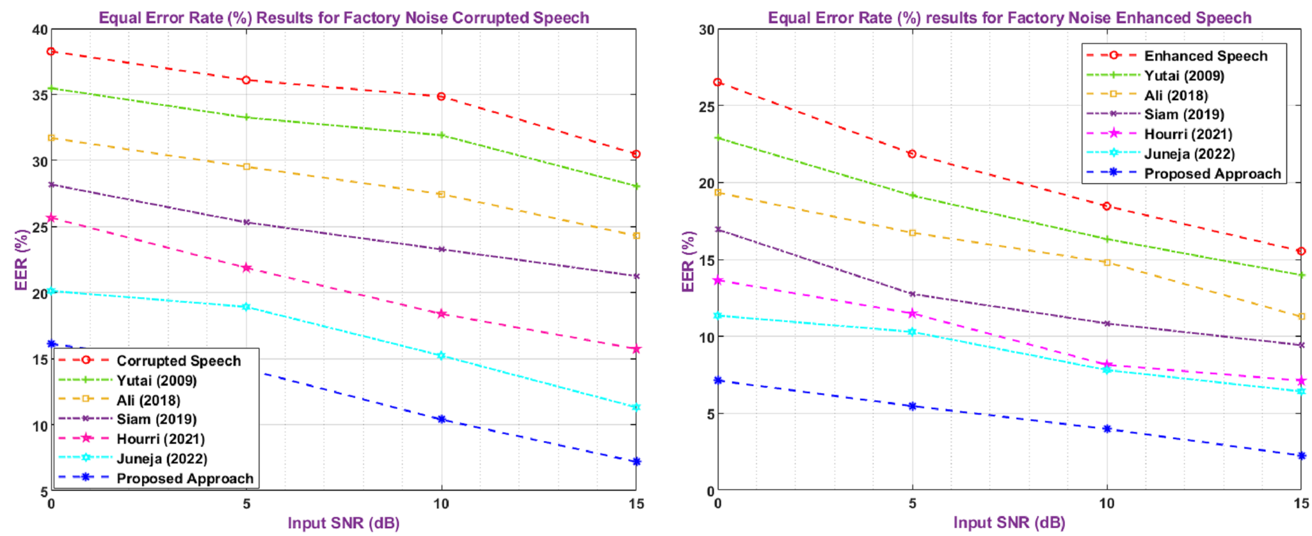

Fig. 5 Speaker verification results (%EER) with MFCC, NPF, and PCC as the input features (factory noise)

15 dB SNR for noise-corrupted speech. For the enhanced for of speech utterance, the FAR achieved is 0.43205 and FRR obtained is 0.11217 at 15 dB SNR level condition. Under all SNR conditions and noise variants, the proposed approach outperforms the existing techniques.

6 Conclusion and future scope

Noise in speech data frequently misrepresents the speaker-dependent features present, complicating speaker identification and verification approaches. Because MFCC is not very resistant to audio degradation processes as a speech classification process, the speaker recognition performance of methods that depend exclusively on MFCC attributes will struggle in the presence of noise encountered in real-time scenarios. Conversely, as demonstrated by the experimental observations, the proposed CNN classifier with the input features of MFCC, NPF, and PNCC is robust to a wide spectrum of audio damages. In terms of identification accuracy, equal error rate, false acceptance rate, and false rejection rate, the proposed technique significantly outperformed all standard procedures by a significant margin.

The future of voice enhancement and speaker recognition is defined by the incorporation of sophisticated signal processing technologies, such as deep learning architectures, as well as the investigation of multimodal approaches that combine auditory and visual information for increased accuracy. Adaptive systems capable of dynamically responding to ambient elements and user context are planned, coupled with attempts to improve resilience against numerous sources of variability such as accents, noise, and channel distortions. As these technologies become more widely used, there will be a greater emphasis on privacy and security concerns. Real-time

Table 2 False acceptance rate (FAR) and false rejection rate (FRR) results (factory noise)

Input SNR	Noise-corrupted speech		Yutai (2009)		Ali (2018)		Siam (2019)		Hourri (2021)		Juneja (2022)		Proposed approach	
	FAR	FRR	FAR	FRR	FAR	FRR	FAR	FRR	FAR	FRR	FAR	FRR	FAR	FRR
0 dB	0.05466	0.02023	0.10472	0.04062	0.21746	0.09275	0.23648	0.09745	0.25164	0.10582	0.28371	0.10824	**0.31547**	**0.11026**
5 dB	0.08257	0.02059	0.13492	0.05049	0.25813	0.09857	0.26318	0.09869	0.27341	0.10713	0.29532	0.10965	**0.32721**	**0.11038**
10 dB	0.10251	0.04075	0.16839	0.05097	0.27148	0.10265	0.28146	0.10372	0.29843	0.10915	0.31469	0.10986	**0.34163**	**0.11045**
15 dB	0.13271	0.06094	0.19527	0.07094	0.28631	0.10543	0.29514	0.10693	0.31467	0.10958	0.32168	0.10995	**0.35942**	**0.11096**
	Enhanced speech		Yutai (2009)		Ali (2018)		Siam (2019)		Hourri (2021)		Juneja (2022)		Proposed approach	
	FAR	FRR	FAR	FRR	FAR	FRR	FAR	FRR	FAR	FRR	FAR	FRR	FAR	FRR
0 dB	0.19531	0.08362	0.22642	0.10571	0.29146	0.10984	0.29837	0.10997	0.32517	0.11059	0.33925	0.11073	**0.38142**	**0.11162**
5 dB	0.20472	0.08954	0.23752	0.10862	0.31572	0.11031	0.31984	0.11047	0.33194	0.11078	0.34719	0.11086	**0.39263**	**0.11184**
10 dB	0.23158	0.09163	0.25961	0.10958	0.32174	0.11053	0.33132	0.11068	0.34912	0.11082	0.35928	0.11092	**0.41603**	**0.11205**
15 dB	0.25837	0.10436	0.27136	0.10974	0.34281	0.11074	0.34963	0.11085	0.35871	0.11096	0.36801	0.11145	**0.43205**	**0.11217**

The text reflected in bold specify the outcomes of our proposed approach

applications in a variety of fields, including healthcare, automotive, security, and customer service, will push the development of efficient algorithms and hardware implementations, allowing for seamless integration into common devices and systems.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability The datasets generated during and/or analysed during the current study are available in the [ITU-T Test Signals for Telecommunication Systems] repository [https://www.itu.int/net/itu-t/sigdb/genaudio/Pseries.htm].

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Jayanna HS, Prasanna SM (2009) Analysis, feature extraction, modeling and testing techniques for speaker recognition. IETE Tech Rev 26(3):181–190. https://doi.org/10.4103/0256-4602.50702
2. Singh N, Khan RA, Shree R (2012) MFCC and prosodic feature extraction techniques: a comparative study. Int J Comput Appl 54(1):9–13
3. Hasan M R, Jamil M, Rabbani MG, Rahman MS (2004) Speaker identification using Mel frequency cepstral coefficients. In: ICECE international conference on electrical & computer engineering, December 2004, pp 565–568
4. Krishnamurthy N, Hansen JH (2009) Babble noise: modeling, analysis, and applications. IEEE Trans Audio Speech Lang Process 17(7):1394–1407. https://doi.org/10.1109/TASL.2009.2015084
5. Yutai W, Bo L, Xiaoqing J et al (2009) Speaker recognition based on dynamic MFCC parameters. In: IEEE international conference on image analysis and signal processing, April 2009. pp 406–409. https://doi.org/10.1109/IASP.2009.5054638
6. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. Digit Signal Process 10(1–3):19–41. https://doi.org/10.1006/dspr.1999.0361
7. Campbell WM, Campbell JP, Reynolds DA et al (2006) Support vector machines for speaker and language recognition. Comput Speech Lang 20(2–3):210–229. https://doi.org/10.1016/j.csl.2005.06.003
8. Campbell WM, Sturim DE, Reynolds DA (2006) Support vector machines using GMM supervectors for speaker verification. IEEE Signal Process Lett 13(5):308–311. https://doi.org/10.1109/LSP.2006.870086
9. Dehak N, Dehak R, Glass JR et al (2010) Cosine similarity scoring without score normalization techniques. In: Odyssey, June 2010. p 15
10. Daqrouq K, Tutunji TA (2015) Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers. Appl Soft Comput 27:231–239. https://doi.org/10.1016/j.asoc.2014.11.016
11. Ajmera PK, Jadhav DV, Holambe RS (2011) Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. Pattern Recognit

44(10–11):2749–2759. https://doi.org/10.1016/j.patcog.2011.04.009
12. Tirumala SS, Shahamiri SR, Garhwal AS, Wang R (2017) Speaker identification features extraction methods: a systematic review. Expert Syst Appl 90:250–271. https://doi.org/10.1016/j.eswa.2017.08.015
13. Jia Y, Chen X, Yu J et al (2021) Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network. Complex Intell Syst 7:1749–1757. https://doi.org/10.1007/s40747-020-00172-1
14. Richardson F, Reynolds D, Dehak N (2015) Deep neural network approaches to speaker and language recognition. IEEE Signal Process Lett 22(10):1671–1675. https://doi.org/10.1109/LSP.2015.2420092
15. Ahmad KS, Thosar AS, Nirmal JH, Pande VS (2015) A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network. In: IEEE eighth international conference on advances in pattern recognition, January 2015. pp 1–6. https://doi.org/10.1109/ICAPR.2015.7050669
16. Soleymanpour M, Marvi H (2017) Text-independent speaker identification based on selection of the most similar feature vectors. Int J Speech Technol 20:99–108. https://doi.org/10.1007/s10772-016-9385-x
17. Liu Z, Wu Z, Li T, Li J, Shen C (2018) GMM and CNN hybrid method for short utterance speaker recognition. IEEE Trans Industr Inform 14(7):3244–3252. https://doi.org/10.1109/TII.2018.2799928
18. Ali H, Tran SN, Benetos E et al (2018) Speaker recognition with hybrid features from a deep belief network. Neural Comput Appl 29:13–19. https://doi.org/10.1007/s00521-016-2501-7
19. Siam AI, El-khobby HA, Elnaby MMA et al (2019) A novel speech enhancement method using Fourier series decomposition and spectral subtraction for robust speaker identification. Wirel Pers Commun 108:1055–1068. https://doi.org/10.1007/s11277-019-06453-4
20. Kenny P (2010) Bayesian speaker verification with, heavy tailed priors. In: Proceedings Odyssey, 2010
21. Taherian H, Wang ZQ, Chang J, Wang D (2020) Robust speaker recognition based on single-channel and multi-channel speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 28:1293–1302. https://doi.org/10.1109/TASLP.2020.2986896
22. El-Moneim SA, Nassar MA, Dessouky MI et al (2020) Text-independent speaker recognition using LSTM-RNN and speech enhancement. Multimedia Tools Appl 79:24013–24028. https://doi.org/10.1007/s11042-019-08293-7
23. Hourri S, Nikolov NS, Kharroubi J (2021) Convolutional neural network vectors for speaker recognition. Int J Speech Technol 24:389–400. https://doi.org/10.1007/s10772-021-09795-2
24. Juneja K (2022) Two-level noise robust and block featured PNN model for speaker recognition in real environment. Wirel Pers Commun 125(4):3741–3771. https://doi.org/10.1007/s11277-022-09734-7
25. Hamidi M, Zealouk O, Satori H et al (2023) COVID-19 assessment using HMM cough recognition system. Int J Inf Technol 15(1):193–201. https://doi.org/10.1007/s41870-022-01120-7
26. Al-Shakarchy ND, Obayes HK, Abdullah ZN (2023) Person identification based on voice biometric using deep neural network. Int J Inf Technol 15(2):789–795. https://doi.org/10.1007/s41870-022-01142-1
27. Radha K, Bansal M (2023) Closed-set automatic speaker identification using multi-scale recurrent networks in non-native children. Int J Inf Technol 15(3):1375–1385. https://doi.org/10.1007/s41870-023-01224-8
28. Chelali FZ (2023) Bimodal fusion of visual and speech data for audiovisual speaker recognition in noisy environment. Int J Inf Technol. https://doi.org/10.1007/s41870-023-01291-x
29. Nakagawa S, Wang L, Ohtsuka S (2011) Speaker identification and verification by combining MFCC and phase information. IEEE Trans Audio Speech Lang Process 20(4):1085–1095. https://doi.org/10.1109/TASL.2011.2172422
30. Wu Z, Chng ES, Li H (2012) Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: Thirteenth annual conference of the international speech communication association, 2012
31. ITU-T P-series recommendations. https://www.itu.int/net/itu-t/sigdb/genaudio/Pseries.htm. Accessed 26 July 2020
32. Gibiansky A, Arik S, Diamos G et al (2017) Deep voice 2: multi-speaker neural text-to-speech. Adv Neural Inf Process 30
33. Nisa R, Showkat H, Baba A (2023) The speech signal enhancement approach with multiple sub-frames analysis for complex magnitude and phase spectrum recompense. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2023.120746
34. Paliwal K, Wójcicki K (2008) Effect of analysis window duration on speech intelligibility. IEEE Signal Process Lett 15:785–788. https://doi.org/10.1109/LSP.2008.2005755
35. Varga A, Steeneken HJ (1993) Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun 12(3):247–251. https://doi.org/10.1016/0167-6393(93)90095-3