applied sciences

MDPI

Article

Enhanced Multiple Speakers' Separation and Identification for VOIP Applications Using Deep Learning

Amira A. Mohamed [1,2,*], Amira Eltokhy [3] and Abdelhalim A. Zekry [2]

[1] Department of Electronics Engineering and Communications, Faculty of Engineering, Badr University in Cairo (BUC), Cairo 11829, Egypt
[2] Department of Electronics and Electrical Communications, Faculty of Engineering, Ain Shams University, Cairo 11517, Egypt
[3] Rapid Bio-Labs, 10412 Tallinn, Estonia
* Correspondence: amira.ahmed@buc.edu.eg

Abstract: Institutions have been adopting work/study-from-home programs since the pandemic began. They primarily utilise Voice over Internet Protocol (VoIP) software to perform online meetings. This research introduces a new method to enhance VoIP calls experience using deep learning. In this paper, integration between two existing techniques, Speaker Separation and Speaker Identification (SSI), is performed using deep learning methods with effective results as introduced by state-of-the-art research. This integration is applied to VoIP system application. The voice signal is introduced to the speaker separation and identification system to be separated; then, the "main speaker voice" is identified and verified rather than any other human or non-human voices around the main speaker. Then, only this main speaker voice is sent over IP to continue the call process. Currently, the online call system depends on noise cancellation and call quality enhancement. However, this does not address multiple human voices over the call. Filters used in the call process only remove the noise and the interference (de-noising speech) from the speech signal. The presented system is tested with up to four mixed human voices. This system separates only the main speaker voice and processes it prior to the transmission over VoIP call. This paper illustrates the algorithm technologies integration using DNN, and voice signal processing advantages and challenges, in addition to the importance of computing power for real-time applications.

Keywords: speaker separation; speaker identification; deep learning; VoIP

Citation: Mohamed, A.A.; Eltokhy, A.; Zekry, A.A. Enhanced Multiple Speakers' Separation and Identification for VOIP Applications Using Deep Learning. *Appl. Sci.* **2023**, *13*, 4261. https://doi.org/10.3390/app13074261

Academic Editors: Yoshinobu Kajikawa and Cheng-Yuan Chang

Received: 18 February 2023
Revised: 20 March 2023
Accepted: 24 March 2023
Published: 28 March 2023

1. Introduction

Voice over Internet Protocol (VoIP) is the internet-based delivery of telephony services enabled by a combination of communication technologies, methods, protocols, and transmission techniques [1]. Instead of analogue phone lines, VoIP allows audio calls to be delivered over IP networks such as the internet. Since the COVID-19 pandemic began, VoIP applications have become an essential part of our daily lives. As a result of audio calls, video calls, and conferencing, the user base of VoIP applications has grown. Whether at work or school, almost everything is done through VoIP applications [2].

Currently, the online call system depends on noise cancellation and call quality enhancement. However, this does not address multiple human voices over the call. Filters used in the call process only remove the noise and the interference from speech signal. However, by adding the Speaker Separation and Identification (SSI) system, it allows sending only the main target speaker rather than any human voices that may exist around the main speaker.

The former studies of Speaker Recognition focus on authentication and security applications, as there has long been a demand to be able to recognise someone solely based on their speech [3]. Furthermore, in the context of speech separation, the requirements of a real-time processing system are even higher and more sensitive. Hearing aids are

an example of speech separation implementation that requires real-time processing [4,5]. Conversely to these applications used in speaker recognition and speaker separation, this research uses both speaker recognition and separation in a different application, which is VoIP.

Figure 1 illustrates that humans can focus on the voice produced by a single speaker in a crowded and noisy environment where they can perform speech–non-speech separation (de-noising) and speaker separation (multi-speaker talking separation) simultaneously [6]. This simple task for humans has proven extremely challenging to mimic in speech processing systems. As the demand for a system that can perform real-time processing is growing, this research focuses on using both speaker recognition and real-time speech separation models to improve the online call process or (VoIP). Our contributions are: (i) A novel SSI which is an integration of deep learning audio separation and identification models, (ii) A separation test up to four speakers at a time, (iii) Involving this model in a VoIP functional system.

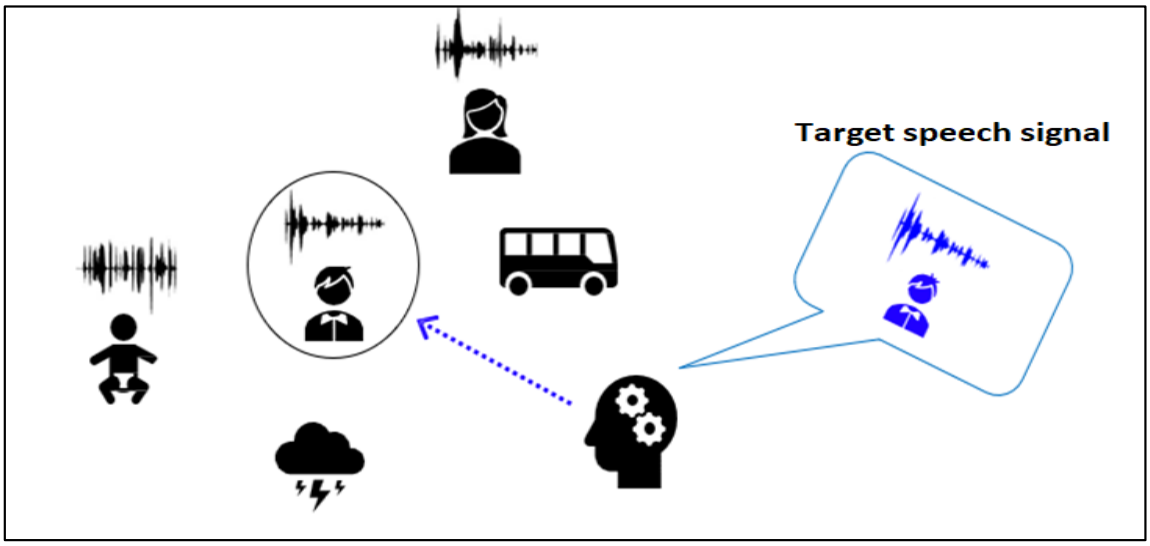

Figure 1. Humans can focus on the target speech signal in a crowded and noisy environment [6].

This paper is organised as follows. Section 2 presents the literature review of speaker identification and speaker separation methods. In Section 3, the proposed model description and the main block diagram of our study are discussed. In Section 4, the setup for conducting our experiment is described. Section 5 presents and discusses the results. Section 6 contains the conclusion.

2. Literature Review

The majority of speaker identification techniques used in the literature before deep learning models were based on i-vector techniques [7–9].

These techniques look for patterns in audio signals and categorise them using methods such as the Mel Frequency Cepstral Coefficient (MFCC), which is essential for identifying speakers. Alternatives to MFCC are provided, including Linear Predictor Coefficient (LPC), Line Spectral Frequency (LPF), Rhythm, Chroma Factor, Turbulence, Spectrum Sub-band Centroid (SSC), and other feature lists. The most frequently employed model for training on our data is the Gaussian Mixture Model (GMM). Other related models, such as the Model of Hidden Markov Chains, can also be used for training (HMM). Recently, the majority of the model training phase for a speaker identification project was carried out using Artificial Neural Networks (ANN).

When compared to many traditional approaches, deep learning algorithms have significantly improved the state of the speech separation topic in recent years [10–15]. A basic neural network speech separation approach begins by applying the short-time

Fourier transform (STFT) to the combined sound to obtain a representation of time and frequency (T-F). After that, inverse STFT is used to synthesise the source waveforms from the T-F bins associated with each source. This concept raises several concerns. First, it is debatable whether Fourier decomposition is the best way to convert the signal for speech separation. Secondly, the signal's amplitude and phase must be taken into account by the separation algorithm because STFT converts the signal into a complex domain. Because altering the phase is difficult, the bulk of proposed approaches merely adjust the STFT's magnitude by generating a time-frequency mask for each source and synthesising using the masked magnitude spectrogram with the mixture's initial phase. As a result, separation performance is constrained above. Even though numerous techniques, such the phase-sensitive mask [16] and complex ratio mask [17], have been developed to leverage phase information to create the masks, because the reconstruction is inaccurate, the upper bound remains. Furthermore, for effective speech separation in the STFT domain, high-frequency resolution is required, resulting in relatively long time windows, which are typically greater than 32 ms for speech [12–14] and greater than 90 ms for music separation [18]. Such systems cannot be used in situations where extremely low latency is required, such as in hearable devices or telecommunication systems because the minimal latency of the system is constrained by the size of the STFT time window. Directly modelling the signal in the time-domain is a straightforward technique to get around these challenges. The method in [19] indicated recent success in tasks including voice recognition, speech synthesis, and speech improvement [20–24].

The current leading approach [19], which is based on an overcomplete set of linear filters, divides the filter outputs at each time step using a mask for two speakers or a multiplexer for more speakers. The audio is then rebuilt using these incomplete representations. Because the order of the speakers is assumed to be random, one uses a permutation invariant loss during training so that the permutation that minimizes the loss is taken into account (it is difficult to sort sounds). This masking-based approach has certain limitations because as the number of speakers increases, the mask must extract and suppress more information from the representation, making it more difficult to deal with the partial representations stated before. Thus, a mask-free approach that was introduced in [25] is used. This technique uses a series of Recurrent Neural Networks (RNNs) to process the audio and it is advantageous to assess the error following each RNN in order to produce a compound loss that represents the quality of the reconstruction after each layer. RNNs have two directions. A particular kind of residual connection, in which two RNNs operate simultaneously, is used to build each RNN block. The bypass-connected layer input and the element-wise multiplication of the two RNNs are combined to create the output of each layer. Unlike isolating known sources [26], the outputs in this scenario are permutation invariant; thus, voices can transition across output channels, particularly during transitory silent periods. So, this series RNN-based audio separation model is employed.

3. Model Description

Instead of introducing the mixed audio signal (main speaker and other speakers around him) directly to the VoIP system, a system called Speaker Separation and Identification (SSI) is inserted between the microphone and the VoIP system. Figure 2 illustrates the main process of SSI block. The first steps of processing are converting real-time mixed audio signal into a recording one in format of WAV to be processed in speaker separation system. Then, all separated speakers are fed to the speaker identification system to identify the target speaker. During that time, the main speaker is ready to send over IP to start the call process without any other human voices or noise around him. The proposed model integrated with VoIP system is depicted in Figure 3.

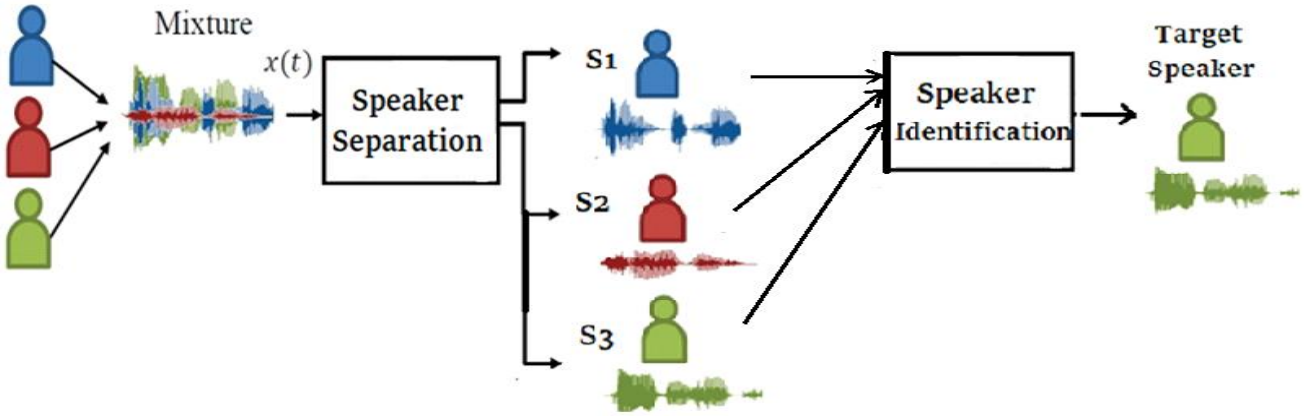

Figure 2. Proposed model integrated with speaker separation and identification for real-time application.

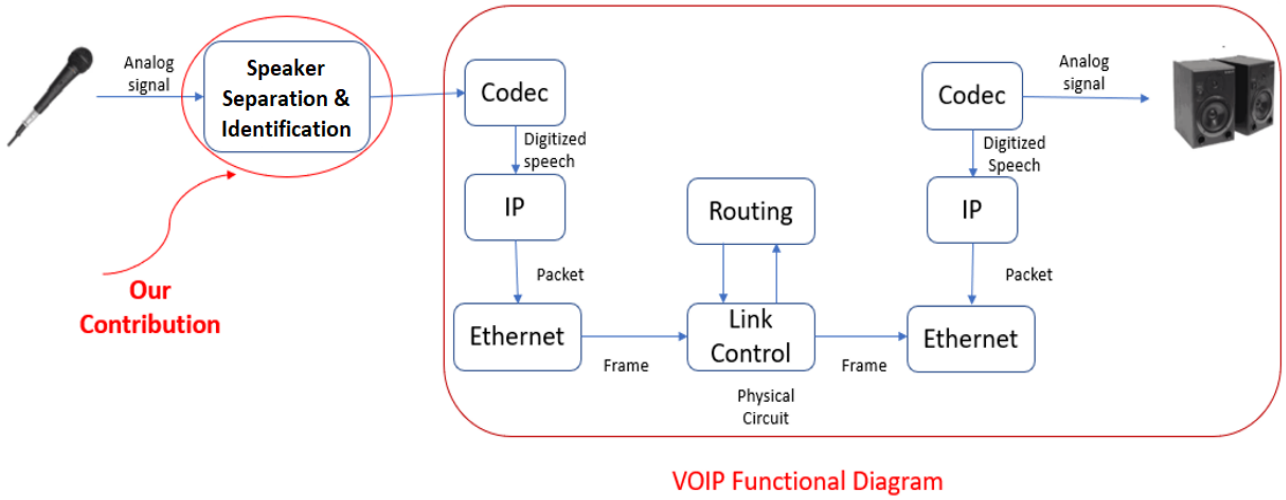

Figure 3. Proposed model integrated with VoIP system.

3.1. Real-Time Audio Pre-Processing

In this part, authors recorded their mixed audios from their microphone and saved it in a NumPy array using the python-sound device module.

This module, combined with the wavio and scipy modules, converts this useful data type for sound processing into WAV format for storage.

The recorded audio files will be used as an input to the separation process module as shown in the next section. The time for these audio files is 7 s for each mix file, with sampling frequency 8 kHz.

3.2. Voice Separation for Multiple Speakers

Real-world voice communication frequently occurs in congested, multi-talker settings. A speech processing system intended to work under such settings must be able to differentiate the speech of distinct talkers. Given a dataset of the main speaker voice and mixed versions of this main voice with other speakers, the system is trained to recognise the speaker voice features using deep learning algorithm. So, at this stage, the system allows the main speaker voice to be separated and identified from other mixed voices.

3.2.1. Speaker Separation Model

Estimating C separate input sources $s_i \epsilon R^T$, where $i \epsilon [1, \ldots, C]$ is the goal of the single-channel source separation problem, given a mixture $x(t) = \sum_{i=1}^{C} s_i(t)$. Since the durations of the input utterances might vary, T, the input length, is not a constant value.

This paper focuses on the supervised setting, where the training set is $S = \{x_i, (s_{i,1} \ldots s_{i,c})\}_{i=1}^{n}$, and the goal is to learn the model to output C estimated channels $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_c)$ in which the scale invariant source-to-noise ratio (SI-SNR) between the estimated and the target utterances is maximized where an unseen mixture x is given. This model was presented in [27] and had the advantage of a lower minimum latency and smaller model size, making it a good choice for both offline and real-time speech separation applications.

3.2.2. Separation Training Objective

We can directly use the source-to-distortion ratio (SDR) as our training objective because the network's output is the waveform of the predicted clean signals. As the training objective, we employ the scale invariant source-to-noise ratio (*SI-SNR*), which is used as the assessment measure in [12,14], instead of the usual SDR. Following is a definition of the *SI-SNR*:

$$s_{target} = \frac{\langle \hat{s}, s \rangle s}{\|s\|^2} \tag{1}$$

$$e_{noise} = \hat{s} - s_{target} \tag{2}$$

$$SI - SNR = 10\, log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \tag{3}$$

where $\hat{s}\ \epsilon\ R^{1\times t}$ is the estimated source and $s\ \epsilon\ R^{1\times t}$ is the target clean source.

The length of the signals is denoted by t. To ensure scale-invariance, $\hat{s}$ and s are both normalized to have zero-mean. The source permutation problem [12–14] is solved using permutation invariant training (PIT) [13], which is used during training. Speaker Identification Model.

3.3. Speaker Identification Process

3.3.1. Speaker Identification Model

The employed methodology first involves model training by GMM, followed by feature extraction through MFCC. Figure 4 provides a visual explanation of this methodology. It is calculated using 20 MFCCs and 20 Delta-MFCCs. There are thus 40 features available in total. Under the feature extraction module, a specially defined function calculated the delta MFCC. One crucial feature extraction method for speaker identification is the MFCC. Finding a set of utterance characteristics that are acoustically correlated to the speech signal, i.e., parameters that can be calculated or roughly estimated is the aim of feature extraction. These variables are referred to as features [28].

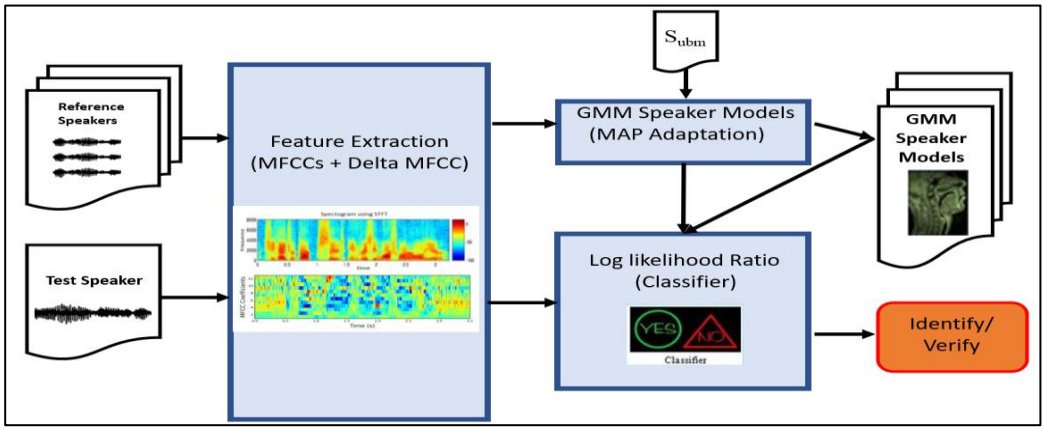

Figure 4. Algorithm flow chart GMM speaker model.

Using a specific feature vector that was extracted from each speaker, the modelling technique aims to produce models for each speaker. By simulating the distributions of the feature vectors, it reduces the amount of feature data. The dependent and independent speakers constitute the two parts of the speaker reorganisation.

The computer should ignore the speaker-specific characteristics of the speech signal when using the speech reorganisation technique in speaker independent mode, and instead, it extracts the desired message. Alternatively, if the speech reorganisation machine is operating in speaker-dependent mode, it should extract speaker characteristics from the acoustic signal.

The primary goal of speaker identification is speech comparison. During the model training phase, the log-likelihood for each gmm model of each speaker was calculated. It was saved in a separate folder as a database. This data dictionary is used to match the gmm file of a 1: N speaker. The speaker with the highest score is chosen and identified.

3.3.2. Identification Training Objective

False Rejection Rate (FRR) and False Acceptancy Rate (FAR) are two factors that heavily influence how well a speaker identification system performs [29]. FRR occurs when the target speaker is mistakenly identified as a non-target speaker. FAR refers to an error made when identifying a non-target speaker as the intended speaker. In a closed set speaker identification system, however, the top-N correctness [30] and accurate recognition rate (accuracy rate) are often employed to assess the system's effectiveness. The speech could be identified accurately after being matched with the appropriate speaker from the target set and is called the recognition rate. The speech that must be recognised is often identified as coming from the speaker who sounds the most like the target speaker set; the top-1 recognition accuracy rate is another name for this recognition accuracy ratio. The top-N recognition accuracy rate is another evaluation technique: if the right speaker is among the N recognition speakers with the highest degree of similarity. Thus, it is determined that the recognition result is accurate [31].

4. Experiments

4.1. Dataset

4.1.1. For Separation

The used separation model's functioning and accuracy are tested on samples of the downloaded dataset, the WSJ0-2mix dataset. However, to check its reliability, correctness, and the ability to integrate the real-time voice recordings with the separation module, we prepared a dataset ourselves. It was a voice recording of multiple speakers. Some of the speakers were chosen to be relatives such as family members who shared nearly similar genetics and features which make them harder to separate. The rest are random speakers such as friends. The combination between relative and non-relative is done to avoid data bias results.

The system is tested on the two-speaker speech separation issue with a custom dataset which has been prepared by the authors. This dataset consisted of four speakers, two males and two females. The dataset includes about four hours of speech divided into 1800 samples for each speaker, with 53.9 h of training data and 200 samples for validation and 200 samples for testing data. The mixtures are created by randomly mixing the two speakers' utterances. The original waveforms generated with 44.1 kHz. Then, to reduce the computational cost, these waveforms are downsampled to 8 kHz. We further expand our dataset to three and four speakers with 55.2 h and 58.8 h of training, respectively.

4.1.2. For Identification

We performed this identification on the separated speaker's dataset output from the previous separation block. We achieved an accuracy of 96% regarding the speaker's identification on this dataset. For training data, four speakers each accompanied 1800 voice samples. In addition, 200 voice samples were collected for testing purposes. As a result,

there were a total of 2000 voice sample datasets. Each voice sample lasted about 7 s. So, we had to tune parameters in mfcc. We adjusted nfft values from 512 to 2000 to 1800 to finally, 1500. Figure 5 indicates the state transition diagram for speaker identification.

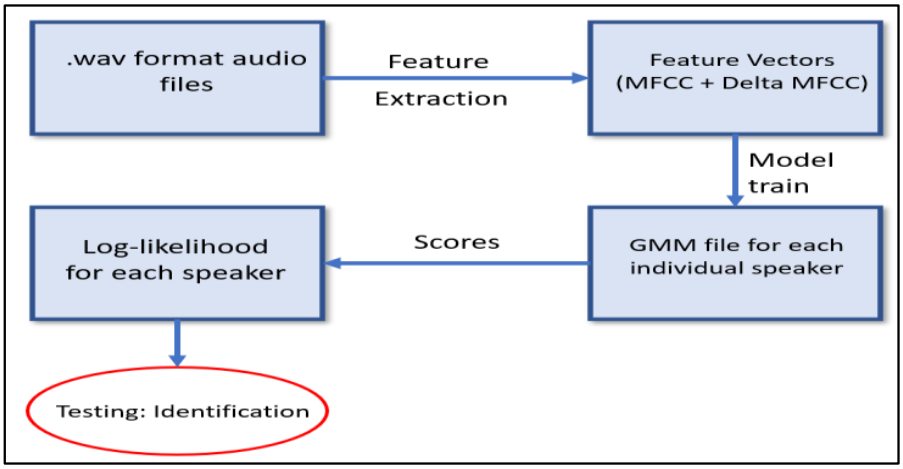

Figure 5. State transition diagram [28].

4.2. Network Configuration

4.2.1. For Separation

Based on the validation set, we choose hyper parameters. The input kernel size L was 8 and the preliminary convolutional layer had 128 filters. A seven-second audio fragment sampled at 8 kHz is used. The design used of b = 6 MULCAT blocks, with 128 neurons in each LSTM layer. We extract the STFT for the speaker model using a 20 ms window, a stride of 10 ms, and a Hamming window.

4.2.2. For Identification

The focus here is on implementing MFCC and GMM concurrently. The main feature was MFCC with tuned parameters with delta-MFCC as a secondary function. In addition, to train our model, we employed GMM with extra fine-tuned parameters. We performed this identification on the separated speaker's output dataset from the previous separation block. We achieved an excellent result on these datasets, with 96% accuracy on the self-prepared dataset. Combining MFCC and GMM in tandem showed great accuracy in identification results in performing speaker recognition tasks.

4.3. Evaluation Metrics

4.3.1. For Separation

The scale invariant signal-to-noise ratio improvement (*SI-SNRi*) score [25] is applied on the test set to evaluate the employed model. This score is calculated as follows:

$$SI\_SNRi(s, \hat{s}, x) = \frac{1}{C} \sum_{i=1}^{c} SI\_SNR(s\_i, \hat{s}_i) - SI\_SNR(s_i, x) \quad (4)$$

We defined s_i, $\hat{s}_i$, x as the ground truth (clean) signal, the estimated audio signal, and the mixed audio signal, respectively. C is the number of speakers. The SI-SNR $(s_i, \hat{s}_i)$ is defined as the estimated scale invariant signal-to-noise ratio while the SI-SNR (s_i, x) belongs to the clean signal.

4.3.2. For Identification

The evaluation of the used identification model is applied on the test set by dividing the total number of voices that have been successfully recognised (*TNSV*) by the total number of voices that have been tested (*TNTV*), which represents the accuracy as follows:

$$Accuracy = \frac{TNSV}{TNTV} \quad (5)$$

5. Results and Discussion

According to different evaluation metrics for both separation and identification as illustrated in Section 3.3, the proposed system successfully separated and identified the main speaker using deep learning. The results showed SNR training curves for various numbers of speakers N = 2, 3, 4, as shown in Figure 6. The used model trained faster for a smaller number of speakers, as reported in Table 1. It demonstrated the performance of the used separation model as a function of the number of mixed speakers compared to the WSJ0-2mix, WSJ0-3mix and WSJ0-3mix dataset. In [25], all reported numbers are the scale invariant signal-to-noise-ratio improvement (SI-SNRi) over the input mixture. It is noted that the SI-SNRi of our self-made experiment for two speakers is less than that of WSJ0-2mix data. One possible explanation for this can be considered, as the relative speakers had more common features which made it harder to be separated than random speakers in the WSJ0-2mix dataset. While the other SI-SNRi, in case of three or four speakers, gave us better results. Figure 7 depicts the SI-SNRi curves for various mixed numbers of speakers. The SI-SNRi increased in parallel with the increasing number of epochs in training, which indicated a better performance of the used model.

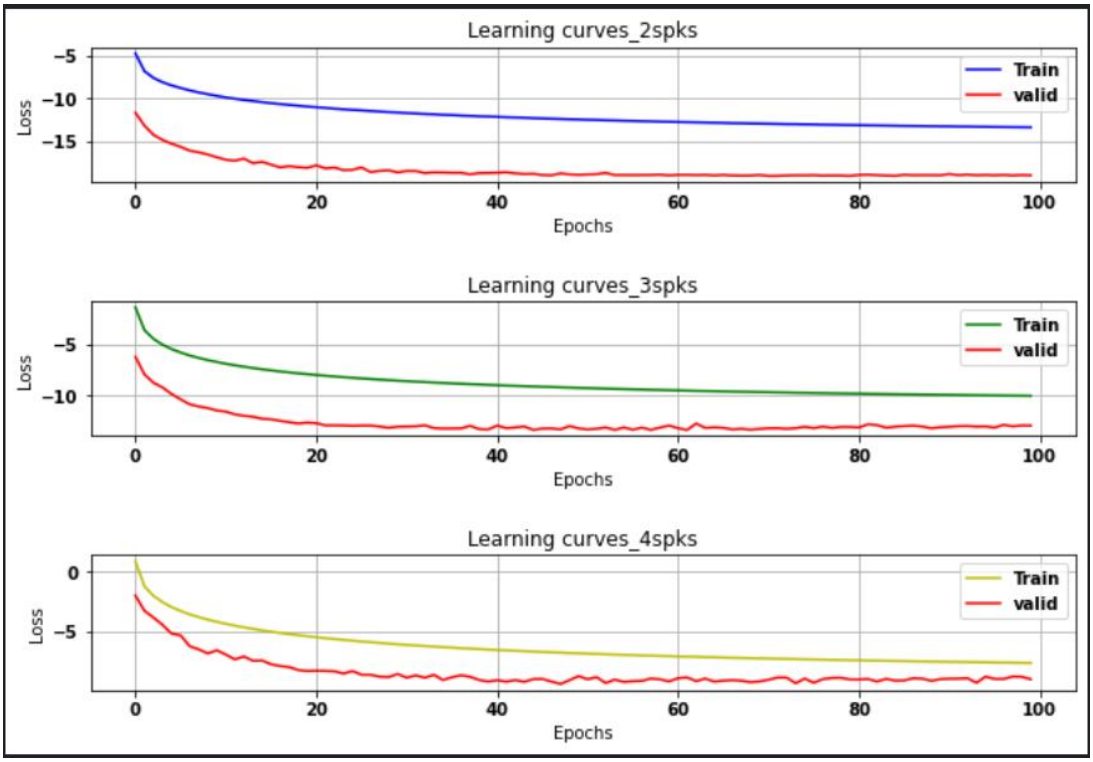

Figure 6. Training curves of our model for various numbers of speakers N = 2, 3, 4.

Table 1. SI-SNRi for different numbers of mixed speakers.

	2 spk	3 spk	4 spk
Self-Made	18.4168	17.5052	15.6260
WSJ0-mix	20.12	16.85	12.88

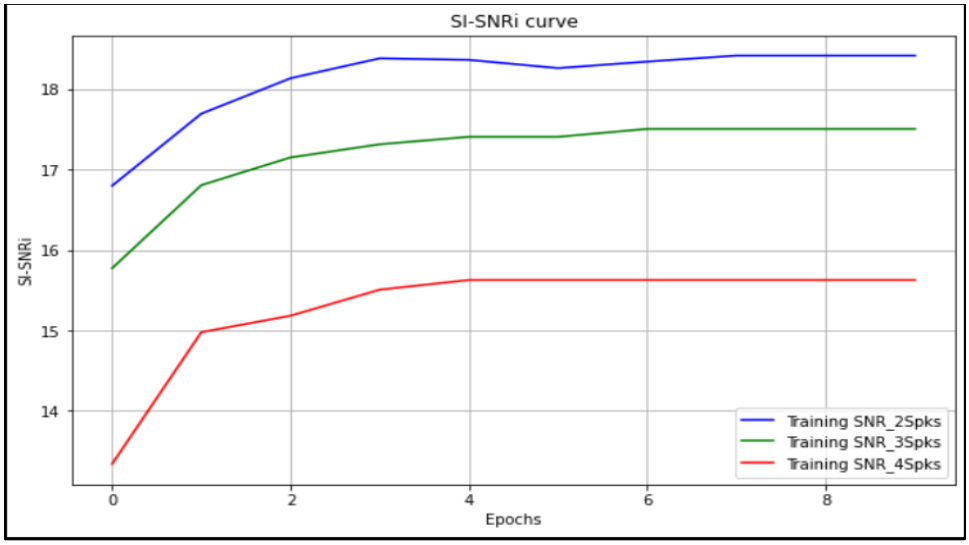

Figure 7. SI-SNRi curves for various mixed numbers of speakers.

On our self-made datasets, speaker identification was carried out successfully and with excellent results, where the accuracy was 96%. Thus, the MFCC-GMM model achieves satisfying performance. Hence, the separated and identified target speaker is ready to be sent alone over a VoIP system without any noise or other speakers attached to it.

The processing time for each block is calculated as it is considered a critical factor for real-time applications. Table 2 shows the process time (testing time) for separation and identification processes separately, then the overall time needed before sending the target speaker's voice over the call.

Table 2. Process time for different numbers of mixed speakers.

	Separation Time	Identification Time	Total Processing Time	Estimated Processing Time after GPU
2 Spk	6.60 s	1.03 s	6.63 s	1.71 s
3 Spk	6.53 s	1.04 s	7.57 s	1.96 s
4 Spk	6.782 s	1.06 s	7.743 s	2.00 s

It is noted that, while increasing the number of mixed speakers, the time needed for separation is increased and so is the overall time. These results of processing time are depicted using only one GPU. Evidently, the processing time can be reduced significantly when using multi-GPU support systems which are suitable for speeding up signal-processing and real-time applications, as explained in [32,33]. The geometric mean of all speedups tested across all datasets for that configuration is the acceleration of a given primitive using a specified number of GPUs. The majority of primitives scale well from one to six GPUs, with various datasets for breadth-first search (BFS), single-source shortest path (SSSP), connected components (CC), betweenness centrality (BC), and PageRank (PR) being 2.63, 2.57, 2.00, 1.96, and 3.86 times faster using six (K40) GPUs as illustrated in detail in [32].

6. Conclusions

In this paper, we presented a speaker separation and identification system using deep learning that is integrated on the VoIP call process system to enhance call experiences and reduce the multiple speaker noise. The main aim is to send only the targeted speaker speech over the VoIP calls, unlike previous work, in which enhancing VoIP calls concentrates on reducing, cancelling, or reducing noise. This system provides a practical solution to a new-reality problem using the proposed SSI block which separates and then identifies the main speaker speech. The system integrated multiple technologies such as deep neural network, STFT, and MFCC-GMM. It was capable of separating up to four speakers with satisfactory signal-to-noise ratio. The paper also presented the main challenges such as processing time and how it can be adjusted based on the used VOIP system. Further studies are recommended to enhance and support this suggested study.

Author Contributions: Conceptualization, A.A.M. and A.E.; methodology, A.A.M.; Software, A.A.M.; validation, A.A.M., A.E. and A.A.Z.; formal analysis, A.A.M.; investigation, A.A.M.; resources, A.A.M.; data curation, A.A.M.; writing—original draft preparation, A.A.M.; writing—review and editing, A.E.; Visualization and supervision, A.E. and A.A.Z.; project administration, A.E.; funding acquisition, A.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research receives no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The authors declare that all relevant evaluation data are available in the figures and tables within the article. The generated and analyzed datasets are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mahmoud, M.A. Assessment and Improvement of the Quality of Voice-over-IP Communications. Master's Thesis, Waterford Institute of Technology, Waterford, Ireland, August 2013.
2. Adhilaksono, B.; Setiawan, B. A study of Voice-over-Internet Protocol quality metrics. *Procedia Comput. Sci.* **2022**, *197*, 377–384. [CrossRef]
3. Campbell, W.M.; Sturim, D.E.; Reynolds, D.A. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Process. Lett.* **2006**, *13*, 308–311. [CrossRef]
4. Wang, D.; Chen, J. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1702–1726. [CrossRef] [PubMed]
5. Wijayakusuma, A.; Gozali, D.R.; Widjaja, A.; Ham, H. Implementation of Real-Time Speech Separation Model Using Time-Domain Audio Separation Network (TasNet) and Dual-Path Recurrent Neural Network (DPRNN). *Procedia Comput. Sci.* **2021**, *179*, 762–772. [CrossRef]
6. Lee, H. Speech Separation, Deep Clustering. Available online: https://blog.csdn.net/qq_45866407/article/details/106878854 (accessed on 22 November 2022).
7. Dehak, N.; Kenny, P.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-End Factor Analysis for Speaker Verification. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 788–798. [CrossRef]
8. Tsujikawa, M.; Nishikawa, T.; Matsui, T. I-vector-based speaker identification with extremely short utterances for both training and testing. In Proceedings of the 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE), Nagoya, Japan, 24–27 October 2017; pp. 1–4. [CrossRef]
9. Travadi, R.; Van Segbroeck, M.; Narayanan, S. Modified-prior i-vector estimation for language identification of short duration utterances. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 3037–3041. [CrossRef]
10. Huang, P.S.; Kim, M.; Hasegawa-Johnson, M.; Smaragdis, P. Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 2136–2147. [CrossRef]
11. Zhang, X.L.; Wang, D. A Deep Ensemble Learning Method for Monaural Speech Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 967–977. [CrossRef] [PubMed]
12. Isik, Y.; Le Roux, J.; Chen, Z.; Watanabe, S.; Hershey, J.R. Single-Channel Multi-Speaker Separation Using Deep Clustering. *Proc. Interspeech 2016* **2016**, 545–549. [CrossRef]

13. Kolbæk, M.; Yu, D.; Tan, Z.H.; Jensen, J. Multitalker Speech Separation with Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1901–1913. [CrossRef]
14. Chen, Z.; Luo, Y.; Mesgarani, N. Deep attractor network for single-microphone speaker separation. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 246–250. [CrossRef]
15. Luo, Y.; Chen, Z.; Mesgarani, N. Speaker-Independent Speech Separation with Deep Attractor Network. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 787–796. [CrossRef]
16. Erdogan, H.; Hershey, J.R.; Watanabe, S.; Le Roux, J. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Queensland, Australia, 19–24 April 2015; pp. 708–712.
17. Williamson, D.S.; Wang, Y.; Wang, D.L. Complex Ratio Masking for Monaural Speech Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *24*, 483–492. [CrossRef] [PubMed]
18. Luo, Y.; Chen, Z.; Hershey, J.R.; Le Roux, J.; Mesgarani, N. Deep clustering and conventional networks for music separation: Stronger together. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 61–65. [CrossRef]
19. Luo, Y.; Mesgarani, N. TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 696–700. [CrossRef]
20. Sainath, T.N.; Weiss, R.J.; Senior, A.; Wilson, K.W.; Vinyals, O. Learning the speech front-end with raw waveform CLDNNs. In Proceedings of the INTERSPEECH, Dresden, Germany, 6–10 September 2015; pp. 1–5.
21. Ghahremani, P.; Manohar, V.; Povey, D.; Khudanpur, S. Acoustic modelling from the signal domain using CNNs. In Proceedings of the INTERSPEECH, San Francisco, CA, USA, 8–12 September 2016; pp. 3434–3438. [CrossRef]
22. Oord, A.V.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. 2016, pp. 1–15. Available online: http://arxiv.org/abs/1609.03499 (accessed on 7 May 2022).
23. Mehri, Y.B.S.; Kumar, K.; Gulrajani, I.; Kumar, R.; Jain, S.; Sotelo, J.; Courville, A. SampleRNN: An unconditional end-to-end neural audio generation model. *arXiv* **2017**, arXiv:1612.07837.
24. Pascual, S.; Bonafonte, A.; Serra, J. SEGAN: Speech Enhancement Generative Adversarial Network. In Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017), Stockholm, Sweden, 20–24 August 2017; pp. 3642–3646. [CrossRef]
25. Nachmani, E.; Adi, Y.; Wolf, L. Voice separation with an unknown number of multiple speakers. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; Volume PartF16814, pp. 7121–7132.
26. Défossez, A.; Usunier, N.; Bottou, L.; Bach, F. Music Source Separation in the Waveform Domain. 2019. Available online: http://arxiv.org/abs/1911.13254 (accessed on 11 May 2022).
27. Luo, Y.; Mesgarani, N. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256–1266. [CrossRef] [PubMed]
28. Singh, B.P.; Jha, S.K.; Khurmi, R.; Yadav, N.K. Live Speaker Identification Using MFCC and Delta-MFCC. *IJSRD—Int. J. Sci. Res. Dev.* **2020**, *8*, 465–470. Available online: https://www.ijsrd.com/articles/IJSRDV8I30362.pdf (accessed on 11 May 2022).
29. Campbell, J.P. Linear Prediction Residual based Short-term Cepstral Features for Replay Attacks Detection. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 1437–1462. [CrossRef]
30. Yong, F.; Xinyuan, C.; Ruifang, J. Evaluation of the deep nonlinear metric learning based speaker identification on the large scale of voiceprint corpus. In Proceedings of the 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China, 17–20 October 2016; pp. 1–4. [CrossRef]
31. Ye, F.; Yang, J. A Deep Neural Network Model for Speaker Identification. *Appl. Sci.* **2021**, *11*, 3603. [CrossRef]
32. Pan, Y.; Wang, Y.; Wu, Y.; Yang, C.; Owens, J.D. Multi-GPU Graph Analytics. In Proceedings of the 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS), Orlando, FL, USA, 29 May–2 June 2017; pp. 479–490. [CrossRef]
33. Schaetz, S.; Uecker, M. A multi-GPU programming library for real-time applications. In *Algorithms and Architectures for Parallel Processing*; Springer Nature: Berlin/Heidelberg, Germany, 2012; Volume 7439, pp. 114–128.