Developing a Low SNR Resistant, Text Independent Speaker Verification System for Intercom Solutions - a Case Study

Szymon Zaporowski
Department of Multimedia Systems, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland *szyzapor@pg.edu.pl*

Franciszek Górski
Department of Multimedia Systems, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland *franciszek.gorski@pg.edu.pl*

Józef Kotus
Department of Multimedia Systems, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland *jozef.kotus@pg.edu.pl*

Abstract

This article presents a case study on the development of a biometric voice verification system for an intercom solution, utilizing the DeepSpeaker neural network architecture. Despite the variety of solutions available in the literature, there is a significant deficiency in the evaluations of text-independent systems under real conditions and with varying distances between the speaker and the microphone. This article aims to bridge this gap. The study explores the impact of different types of parameterizations on network performance, the effects of signal augmentation, and the results obtained under conditions of low Signal-to-Noise Ratio (SNR) and reverberation. The findings indicate a significant need for further research, as they suggest substantial room for improvement.

Keywords: Speech Biometrics, Speech Processing, Speaker Verification, DeepSpeaker

1. Introduction

Speaker verification technology, which intricately analyzes and verifies the identity of individuals based on their vocal characteristics, has become a cornerstone of modern auditory analysis systems. This paper delves into the development and refinement of Speaker Verification Systems (SVS) specifically engineered to operate under challenging acoustic conditions marked by low Signal-to-Noise Ratios (SNR), including significant levels of environmental noise and reverberation. One of the crucial points investigated is the impact of the distance between the speaker and the microphone. This aspect of the research addresses how varying distances can affect the capture and quality of the audio signal, further complicating speaker verification under non-ideal conditions. The primary aim is to enhance the robustness and accuracy of these systems, ensuring reliable identification across varied and non-ideal auditory environments, and providing seamless scalability for integration into devices with limited processing capabilities, such as intercom systems.

The field of speaker recognition overlaps with numerous areas such as signal processing, machine learning, and psychoacoustics, creating a multifaceted set of obstacles mainly due to harsh environmental factors. These difficulties are further intensified by elements like background noise from city environments, echoes within buildings, and various fluctuating interferences that considerably lower the quality of audio and obstruct

reliable speaker verification.

This research integrates signal-processing techniques to mitigate the detrimental effects of noise and reverberation, enhancing the signal clarity and integrity. Concurrently, the study harnesses robust acoustic feature extraction methodologies including Mel-Frequency Cepstral Coefficients (MFCCs), which are crucial for capturing the unique aspects of speaker voices, and Gammatone Frequency Cepstral Coefficients (GFCCs), which more closely mimic the human auditory system's response than traditional methods [1], [3], [33]. Additionally, the research incorporates the use of melspectrograms, which provide a visual representation of the spectral composition of sounds, similar to MFCCs but retaining more spectral detail, making them useful in complex auditory environments [8] .

The research objective of this study is to identify the most effective solution for an intercom speaker verification system. The ideal system should: (1) achieve the lowest possible Equal Error Rate (EER), (2) demonstrate robust performance in noisy and reverberant environments, and (3) require minimal maintenance, utilizing the intercom's computational resources for signal parametrization, (4) should work properly in various distances between speaker and microphone.

The paper is organized as follows: Section 2 outlines the related work in the discussed filed. Section 3 describes the feature extraction methods utilized, employed dataset and machine learning algorithm adapted for robust speaker recognition; Section 4 presents the results; Section 5 discusses the results and performance comparisons of different models; and Section 6 concludes by presenting future research directions and the practical implications of this study.

2. Related Work

The issues concerning speaker verification systems are a prevalent topic within scholarly literature. The current trend indicates a shift towards the development of sophisticated systems primarily based on deep neural networks and architectures such as encoder-decoder frameworks or Time Delay Neural Network (TDNN) models [7], [16], [21], [34], [36]. Older approaches employ Convolutional Neural Networks (CNN) combined with Recurrent Neural Networks (RNN) especially Long-Short Term Memory (LSTM) networks [28], [35].

In most cases where traditional parameterization methods are used, which do not involve the extraction of parameters directly from the audio signal using neural networks, MFCCs and melspectrograms are typically employed [9, 10], [13–15], [24, 25]. Less commonly used parameters include Linear Predictive Coding (LPC) and GFCCs [4], [6], [26], [31].

The literature features numerous examples of speaker-independent verification systems. However, these systems are seldom tested under conditions of significant reverberation or low SNR [2], [5], [12], [25], [29]. This lack of testing in challenging acoustic environments raises questions about the robustness and reliability of these verification systems when deployed in real-world scenarios.

Moreover, there are only isolated studies on the impact of the distance between the speaker and the microphone on the quality of speaker verification. These studies have been conducted primarily for text-dependent models and in Mandarin Chinese [20]. This indicates a significant gap in the literature, as there is no described system that simultaneously: a) operates independently of the text being spoken, b) functions effectively in environments with low SNR and high reverberation, c) utilizes standard parametrization rather than relying on an encoder model, and d) is adapted for use in the Polish language or nearly language independent e) are tested in various distances between speaker and microphone .

This gap in research underscores the need for further investigation into developing more adaptable and resilient speaker verification systems that can operate effectively across different languages and under adverse acoustic conditions. Such advancements would enhance the security and applicability of speaker verification technologies in a wide range of applications, from secure communications to personalized user interfaces.

MOST WIEDZY Downloaded from mostwiedzy.pl

3. Methodology

The intercom solution is comprised of three principal components: an edge device (intercom) functioning as both a data recorder and parameterizer, which includes a Voice Activity Detection (VAD) algorithm, centralized repository and AI based model for speaker verification. The VAD algorithm employs an energy-based method for speech detection, effectively differentiating between speech and non-speech segments by analyzing the energy levels of the audio input. Previously mentioned centralized repository archives registered user biometric samples for subsequent analysis. An artificial intelligence model is tasked with performing comparative analysis between the biometric samples captured by the intercom and the existing samples stored in the database. These samples undergo a comparative process, and if they surpass a predefined similarity threshold, the system validates the user's identity. Inter-component communication within the system is orchestrated through a bespoke Application Programming Interface (API) specifically developed to meet the system's requirements. The architecture of the system and the flow of information are illustrated in Figure 1.

Given the objective of optimizing resource utilization, the study included a comprehensive analysis of the computational complexity of parameterization algorithms and assessed the impact of parameterization types on training efficacy. This necessitated an experimental evaluation of the input data size, training duration, and Virtual Random Access Memory (VRAM) occupancy on graphic processing units.

Subsequently, the analysis focused on comparing the average prediction time for each parameterization method to determine its tangible impact on performance metrics.

In the final selection phase, the top five models for each type of parameterization were validated and selected from several hundred saved checkpoints. These models were then evaluated based on their parameterization quality, with the optimal model being selected based on real-world data collected in a slightly noisy environment using an intercom system. Model assessments were conducted using both the original signal and a signal artificially corrupted with Gaussian noise, achieving a SNR close to 20 dB.

The quality of the model was quantified using the EER, defined by the following equation (1):

$$\boldsymbol{EER} = \boldsymbol{FAR}(\boldsymbol{\theta EER}) = \boldsymbol{FRR}((\boldsymbol{\theta EER}) \tag{1}$$

Where $\boldsymbol{\theta EER}$ is the threshold at which the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR). FAR and is defined by equation (2):

$$\boldsymbol{FAR} = \frac{\boldsymbol{Number\ of\ False\ Acceptances}}{\boldsymbol{Total\ Number\ of\ Authentication\ Attempts}} \tag{2}$$

FRR is described by equation (3):

$$\boldsymbol{FRR} = \frac{\boldsymbol{Number\ of\ False\ Rejections}}{\boldsymbol{Total\ Number\ of\ Authentication\ Attempts}} \tag{3}$$

These metrics were critical in determining the robustness of the models under conditions that simulate typical usage scenarios. The model that demonstrated the best performance based on these criteria was further utilized in subsequent experiments.

It is important to note that all EER metrics were calculated using the round-robin without repetition (each with each without repetition) comparison of voice samples. This approach resulted in approximately 20,000 comparisons for each of the presented experiments.

The method of round-robin without repetition ensures that each voice sample is compared with every other sample exactly once, thereby maximizing the diversity and comprehensiveness of the testing scenario. This rigorous testing approach provides a robust measure of the system’s ability to correctly identify and authenticate speakers under varied conditions, without the bias that might arise from repeated comparisons of the same samples.

MOST WIEDZY Downloaded from mostwiedzy.pl

The substantial number of comparisons, about 20,000 in this context, underscores the thoroughness of the validation process. Such detailed testing is essential for identifying potential weaknesses in speaker verification systems, particularly in terms of their susceptibility to various types of errors under different operational conditions.

Following the selection of the appropriate parameterization, additional training sessions were conducted to find the optimal settings. Adjustments were made to the learning rate and batch size, and the impact of the first mel-frequency band on the quality of the model was examined. Ultimately, two models were presented, primarily differing in the setting of the first mel-band – the first model starts the mel-band according to the standard settings of librosa and uses a batch size of 128. For the second model, the first mel-band starts at 150 Hz. This value was selected based on empirical evaluation of various frequencies ranging from 50 to 250 Hz.

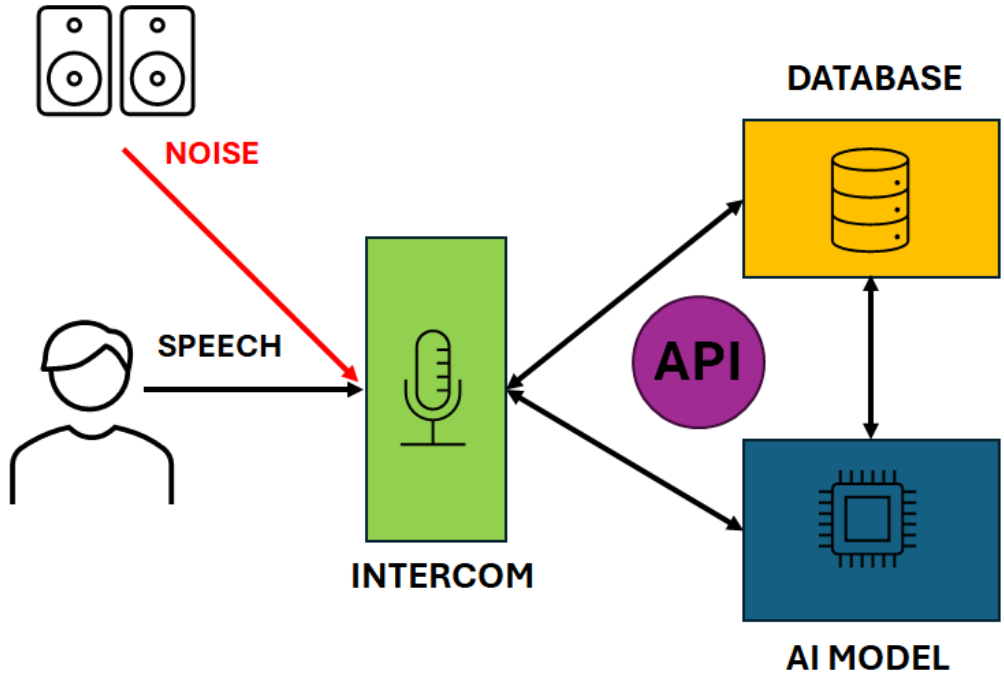

Fig. 1. Diagram of the intercom solution featuring a voice biometrics module (speaker verification) consisting of three modules: intercom, database, and AI model.

The training and validation processes were conducted using a computer equipped with an AMD Ryzen Threadripper 2950X 16-core processor and an Graphical Processing Unit (GPU) - NVIDIA RTX 3090.

3.1. The Dataset

For training purpose, a LibriSpeech dataset was employed [19]. The dataset consists of an extensive collection of about 1,000 hours of English-language audiobooks sourced from the LibriVox project, all available in the public domain. It is widely utilized in the research and development stages of automatic speech recognition (ASR) technologies. The dataset is intentionally compiled to offer a varied and rigorous assortment of speech recordings, relevant for both training and evaluating ASR systems.

LibriSpeech features a diverse array of speakers, showcasing multiple accents and speech styles, thus providing a resource for developing robust ASR systems that can accurately interpret various speech patterns under different conditions. The audio recordings within LibriSpeech are sampled at 16 kHz and include voices from both male and female speakers, spanning a broad spectrum of ages and predominantly North American accents.

The dataset underwent an augmentation process utilizing a suite of advanced audio processing techniques to enhance the robustness and variability of the training samples. This included the injection of Gaussian noise to model stochastic background disturbances, the application of spectral notching to selectively attenuate specific frequency bands, and

MOST WIEDZY Downloaded from mostwiedzy.pl

the convolution of audio signals with impulse responses characteristic of various room acoustics [17, 18]. The MIT McDermott IR dataset, which consists of 271 different impulse responses, was utilized [32]. These impulse responses were meticulously selected to represent rooms with predefined reverberation times, thereby simulating a range of reverberant conditions that speakers might encounter in real-world environments. This augmentation strategy was designed to comprehensively prepare the dataset for effective training of speaker recognition models under diverse and challenging acoustic scenarios.

For validation purposes, a separate dataset was recorded using an intercom system equipped with MEMS-type microphones. Approximately 50 individuals, diverse in gender, were recorded, with each person providing a minimum of four phrases in Polish. The recordings were conducted at various distances from the microphone, ranging from 0.25 to 1 meter, in two rooms with significantly different acoustic characteristics. One of the rooms featured a reverberation time exceeding three seconds. The illustration of the considered rooms is presented in Figure 2.

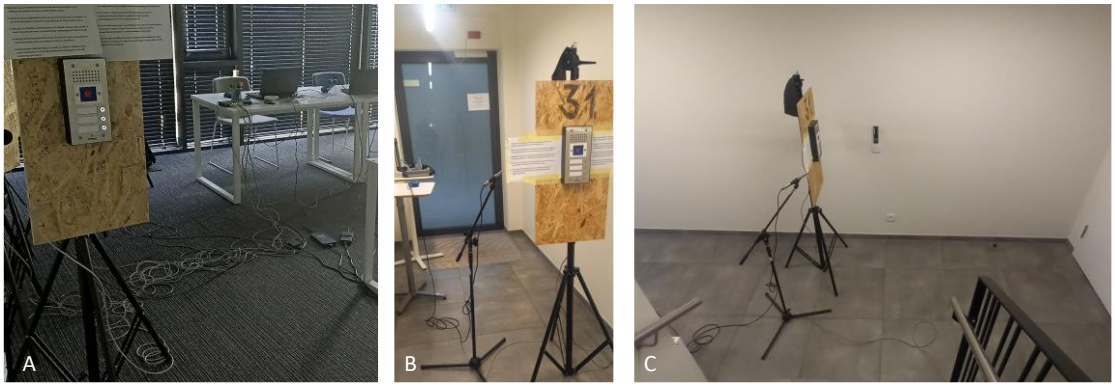

Fig. 2. Illustration of the recording setup in real conditions: A – office room, B and C – reverberant conditions

3.2. Parameterization techniques

For the experiment, three primary methods were utilized for the parametrization of the audio signal: MFCCs, GFCCs, and the Melspectrogram. The selection of these parametrization techniques primarily stemmed from the computational limitations inherent to the intercom responsible for the parametrization process. This constraint necessitated the exclusion of more complex neural network architectures, such as those involving encoder-decoder models or transformer-based architectures, as it was not feasible to execute the encoder component or other sophisticated architectures on the intercom.

Table 1. Settings of parameterization employed in conducted experiments

Parameter	MFCC	GFCC	melspectrogram
Window length	25 ms	25 ms	128 ms
Overlap	10 ms	10 ms	32 ms
Window type	Hamming	Hamming	von Hann
FFT (Fast Fourier Transform) length	512	512	2048
Number of filters	128	64	128
Number of coefficients	40	40	X
Sampling frequency	16 kHz	16 kHz	16 kHz
Number of frames per network input segment	160 frames	160 frames	X

MFCC and GFCC share a similar calculation method, differing primarily in the type of filters used; this was convenient in terms of implementation on an external platform—with a slight modification, two types of parameterizations can be achieved. Based on the

MOST WIEDZY Downloaded from mostwiedzy.pl

literature, GFCCs are expected to perform better than MFCCs in conditions of significant noise, which is a desirable feature for the discussed system. Regarding melspectrograms, they were chosen due to their effective compatibility with CNN. Additionally, unlike MFCCs, melspectrograms retain more detailed information about the spectral and temporal characteristics of the speech, as they do not undergo the final Discrete Cosine Transform (DCT) step.

MFCC and melspectrograms were implemented using the librosa library and GFCC was implemented using Spafe [22, 30]. The utilized settings are presented in the Table 1.

3.3. Employed Architecture

The base architecture chosen for this project is DeepSpeaker, primarily due to the authors' familiarity with this architecture from conducting the other research related to it. Additionally, the system adapts well when trained on one language and tested on another, such as training on Mandarin and testing on English, suggesting its robustness across different linguistic contexts. This was also an important factor due to insufficient training data in Polish to train a relatively large model. Furthermore, the architecture has an implementation in the TensorFlow framework, which continues to be actively developed [23].

DeepSpeaker is an advanced neural network framework meticulously crafted for the purpose of speaker recognition. It leverages a specialized variant of the Residual Network (ResNet) design, integrating auditory processing insights to enhance its capability in distinguishing and identifying various speakers based on their vocal signatures [11].

Central to the architecture of DeepSpeaker is an advanced residual learning strategy, specifically engineered to address the prevalent problem of vanishing gradients that occurs in the training of deep neural networks. Residual Networks are distinguished by their innovative use of "skip connections" which facilitate the direct incorporation of inputs into the outputs at various layers. This distinctive feature supports the training of substantially deeper networks by maintaining an uninterrupted flow of gradients across the network structure, thereby enhancing overall training efficacy.

DeepSpeaker customizes the ResNet framework to suit audio processing needs, particularly focusing on encoding short audio segments into a high-dimensional space. That means the embeddings of the same speaker cluster closer together, markedly distinct from those of different speakers. The primary operational data for the network comprises MFCCs, recognized for their efficacy in encapsulating distinct vocal traits.

During the operational phase of DeepSpeaker, audio inputs are first converted into MFCCs. These coefficients are subsequently fed through multiple residual blocks. Each of residuial blocks which consist of a convolution layer, batch normalization layer, and ReLU activation function. The architecture's skip connections play a pivotal role in counteracting gradient vanishing during this phase.

Following the residual blocks, average pooling is employed to reduce the dimensionality of the feature maps while carefully preserving essential features. A dense layer subsequently maps these condensed features into a tailored embedding space. The dimensions of this embedding space are finely adjusted to meet the demands of the speaker recognition task by reducing the dimensionality of the input vectors.

For training, DeepSpeaker utilizes a triplet loss function, essential for cultivating discriminative features pivotal for speaker verification. This function processes triples of audio samples — an anchor, a positive (another sample from the same speaker), and a negative (from a different speaker). The triplet loss aims to minimize the distance between the anchor and the positive while maximizing that between the anchor and the negative, effectively enhancing the system's ability to differentiate between speakers.

The main modifications introduced to the architecture involve adapting the network inputs to vectors containing solely MFCC and GFCC features, as well as separately for melspectrograms. Another significant change is the elimination of random selection of segments that are used to create the embedding vector based on the input data – the entire vector is considered, rather than just a snippet. This is a crucial adjustment, particularly due to the implementation of melspectrograms. Combined with VAD this modification

MOST WIEDZY Downloaded from mostwiedzy.pl

helps avoid scenarios where the vector describing a person is merely a segment of noise.

4. Results

Table 2 presents the GPU memory utilization, duration of training epochs, and the dimensions of the training data used for Softmax training of the DeepSpeaker model across three parameterization methods: MFCC, GFCC, and melspectrogram. Notably, while MFCC and GFCC consume the same amount of GPU memory and have the same input data dimensions, the epoch duration for GFCC is significantly longer. Melspectrogram utilizes less VRAM and has shorter training epochs, demonstrating potential efficiencies in resource usage.

Table 2. List of GPU VRAM usage, epoch duration, and input data size for Softmax training of DeepSpeaker

Parameterization Method	GPU Memory Usage [MiB]	Training Epoch Duration [s]	Training Data Dimensions (batch, n_rows, n_cols)
MFCC	4405	573	(32, 160, 40)
GFCC	4405	720	(32, 160, 40)
Melspectrogram	3381	382	(32, 128, 32)

Table 3 shows similarity in GPU memory usage across the parameterization methods for Triplet training. However, the Melspectrogram method is not only more memory efficient but also requires substantially less time per epoch compared to MFCC and GFCC, which could translate into faster model training cycles.

Table 3. List of GPU VRAM usage, epoch duration, and input data size for Triplet loss training of DeepSpeaker

Parameterization Method	GPU Memory Usage [MiB]	Training Epoch Duration [s]	Training Data Dimensions (batch, n_rows, n_cols)
MFCC	4405	214	(96, 160, 40)
GFCC	4405	217	(96, 160, 40)
Melspectrogram	3381	130	(96, 128, 32)

In the Table 4 the average prediction times for the DeepSpeaker model using different parameterization methods are listed, with GFCC achieving the fastest prediction time, followed closely by Melspectrogram and MFCC. This suggests that GFCC may offer computational advantages in real-time applications.

Table 4. Average Model Prediction Time with 10 Repetitions for Softmax training of DeepSpeaker

Parmeterization Method	Average Prediction Time [s]
MFCC	0.494
GFCC	0.422
Melspectrogram	0.432

In the Table 6 there is comparison of the prediction times for Triplet trained models, where GFCC and Melspectrogram show significantly better performance over MFCC. This indicates their potential for efficient deployment in time-sensitive scenarios.

Table 5. Average Model Prediction Time with 10 Repetitions for Triplet training of DeepSpeaker

Parmeterization Method	Average Prediction Time [s]
MFCC	0.629
GFCC	0.416
Melspectrogram	0.415

The results presented in the Table 6 demonstrate the effectiveness of each parameterization method under ideal noise-free conditions, with GFCC displaying remarkably lower EERs, suggesting superior model accuracy.

MOST WIEDZY Downloaded from mostwiedzy.pl

Table 6. Average results of validation for different parameterization approaches – no noise scenario

Parameterization name	EER – Softmax [%]	EER – Triplet [%]
MFCC	**6.10**	**5.70**
GFCC	**1.25**	**1.21**
Melspectrogram	**6.30**	**6.20**

In the Table 7 there are given results of verification under noisy conditions. The validation results vary significantly across methods, with GFCC experiencing a drastic increase in EER compared to its performance in a no-noise scenario, while melspectrogram shows a moderate increase. This highlights the challenges of noisy environments in biometric verification.

Table 7. Average results of validation for different parameterization approaches – noise scenario

Parameterization name	EER – Softmax [%]	EER – Triplet [%]
MFCC	**10.20**	**10.40**
GFCC	**24.00**	**21.10**
Melspectrogram	**13.40**	**7.90**

Table 8 presents the Average results of validation for melspectrogram parameterization approach with noise scenario with full noise data augmentation, where all samples were augmented using noise.

This table focuses on the melspectrogram parameterization under a noise scenario with full data augmentation. The results show improved EERs compared to the non-augmented noise scenario, underscoring the effectiveness of data augmentation in enhancing model robustness against noise.

Table 8. Average results of validation for melspectrogram parameterization approach – noise scenario with full noise data augmentation

Parameterization name	EER – Softmax [%]	EER – Triplet [%]
Melspectrogram	**6.00**	**6.30**

Table 9 presents the results of average validation for the mel-spectrogram parameterization approach using real data from Room 1. The table details EER for models tested in real-room scenarios at various distances. The old model is trained with the standard Librosa library settings, augmented by adding Gaussian noise to the audio files. The new model is trained using augmented methods, including adding impulse responses and Gaussian noise to the audio files, and modifying the first mel-band, starting from 150 Hz. The new model generally shows improved accuracy over the old model, particularly at closer distances, which may indicate enhancements in model sensitivity and spatial discrimination.

Table 9. Average results of validation for melspectrogram parameterization approach – Scenario on real data Room 1

Distance from Microphone	Old Model EER [%]	New Model EER [%]
Room1 - 0.5 m	**5**	**3**
Room1 - 1m	**11**	**13**
Room1 - Entire	**11**	**8**

Figure 3 shows average result of the validation process for melspectorgram parameterization approach using real data from Room 2 (with reverberation). These results indicate the performance consistency of the new model across various distances in Room 2, with generally lower EERs than the old model. The results highlight the new model's

MOST WIEDZY Downloaded from mostwiedzy.pl

improved robustness and accuracy in diverse real-world conditions.

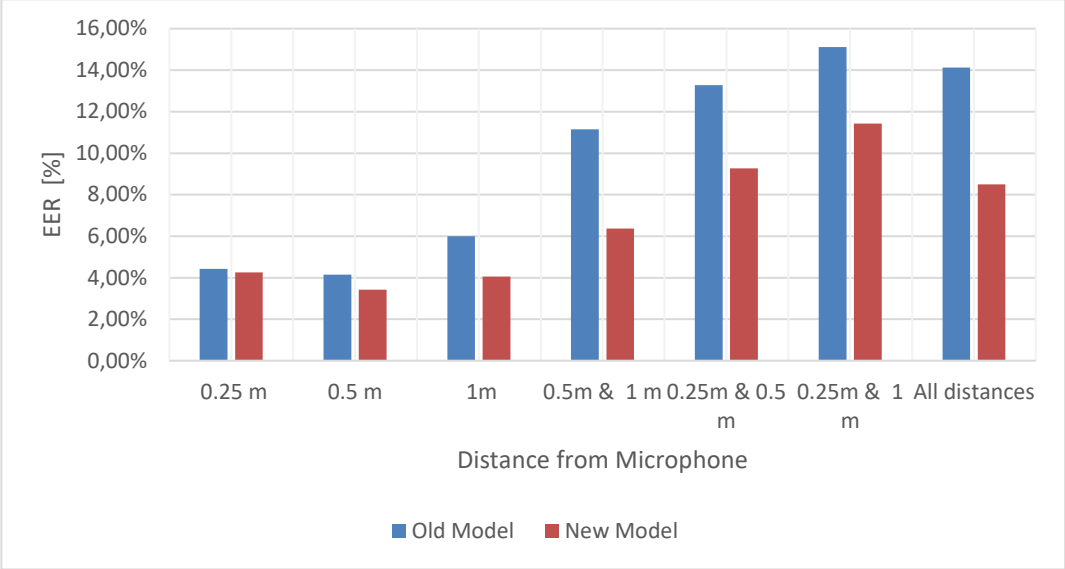

Fig. 3. Average results of validation for melspectrogram parameterization approach – Scenario on real data Room 2

5. Discussion

The results obtained from various experiments on the DeepSpeaker voice biometric system demonstrate that the efficiency and effectiveness of the parameterization methods, as well as the training strategies, have a substantial impact on both the performance metrics and practical deployment considerations of the system.

Firstly, resource utilization analysis revealed significant differences across parameterization methods. The melspectrogram method consistently showed lower GPU VRAM usage and shorter epoch durations as compared to MFCC and GFCC (Tables 2 and 3). This indicates a more efficient allocation of computational resources, which is critical for deployment in resource-constrained environments. Furthermore, the prediction time analysis showed that GFCC generally offered the shortest prediction times, especially with Triplet loss training (Tables 4 and 5), highlighting its potential for applications requiring low latency.

In terms of model accuracy and robustness, the performance under different acoustic conditions varied notably. GFCC demonstrated superior performance under ideal, noise-free conditions but suffered a significant degradation in noisy environments (Tables 6 and 7). This indicates that while GFCC excels in controlled settings, its robustness in adverse conditions is limited. This finding is contrary to the current state of knowledge, necessitating a repetition of the experiment. The repeated experiment yielded the same results, suggesting that either the experiment is poorly designed, or the GFCC parameterization is not capable of handling noisy signals when processed within a DeepSpeaker network consisting of residual blocks of convolutional layers. Conversely, the melspectrogram method, particularly with noise data augmentation, showed improved resilience in noisy conditions, reducing the Equal Error Rates significantly (Table 8). This improvement underscores the importance of integrating realistic noise profiles during the training phase to enhance the model's ability to withstand real-world acoustic disturbances.

The analysis of the model's performance in real-room scenarios revealed that distances from the microphone substantially affect the system's accuracy (Tables 9 and Figure 3). The new model iterations demonstrated improved performance at closer distances, indicating enhancements in model sensitivity and spatial discrimination. Such improvements are vital for the effective deployment in environments like smart homes or security systems where variable user interactions with the device can be expected.

The comparative analysis between Triplet and Softmax training methods showed that Triplet training not only often results in better prediction times but also achieves

MOST WIEDZY Downloaded from mostwiedzy.pl

comparable or superior error rates. This finding suggests that Triplet training may be more suitable for developing efficient and robust biometric systems.

From these findings, further research into adaptive or hybrid parameterization methods that can dynamically adjust to changing acoustic environments could be beneficial. Additionally, investigating more advanced noise augmentation techniques and their integration during model training could lead to further enhancements in robustness. Exploring the impacts of newer deep learning architectures and optimization techniques might also offer additional performance improvements.

Attempting to compare the obtained results with those reported in the literature, it is not possible to directly reference a similar case. In comparing the Equal Error Rate (EER) results for the Hi-Mia dataset and examining the impact of microphone distance from the speaker, it is necessary to consider that the solution presented in that article is text-dependent and that distances greater than 1 meter were also studied. The EER results reported by the authors of that solution range from 3.29% to 4.1% [20]. Solutions based on advanced neural network architectures achieve an EER below 1.5% [7]. However, their evaluation does not occur under as challenging conditions, such as significant noise or reverberation. Considering small distances from the microphone (e.g., 0.25 m), it can be assumed that reverberation does not have such a significant impact on the result. Analyzing the results in this manner, it is evident that they deviate from the state of the art, but the difference is not as substantial. Conducting another comparison with solutions designed to operate in high noise environments, the obtained EER results are around 8% or worse depending on the reverberation time [2], [27] . Comparing the presented results, for small distances, the outcomes are close to the state of the art. However, for combined distances (close and far from the microphone) or greater distances, the results are worse. No studies were found in the literature that validated speaker verification under similar conditions.

Overall, the results from this comprehensive study highlight the complex interplay between model training strategies, parameterization methods, and deployment scenarios. They provide a roadmap for future research aimed at improving the accuracy and robustness of voice biometric systems in both controlled and real-world environments.

6. Conclusion

The comprehensive evaluation of the DeepSpeaker voice biometric system through a series of experiments has provided significant insights into the performance of various parameterization methods under different training and environmental conditions. The findings from this study illuminate both the potential and limitations of current voice biometric technologies and pave the way for future enhancements.

Firstly, the study confirmed that the efficiency of computational resource utilization varies significantly between parameterization methods, with the melspectrogram method demonstrating notable advantages in terms of lower VRAM usage and shorter training times. Such efficiency is crucial for deploying voice biometric systems in resource-constrained settings.

Secondly, while GFCC parameterization showed exceptional performance in noise-free conditions, its susceptibility to performance degradation in noisy environments highlights a critical vulnerability. This contrast underlines the necessity for robustness in practical deployment scenarios, especially in environments with variable acoustic conditions.

Moreover, the application of noise data augmentation techniques, particularly with the melspectrogram method, markedly improved model resilience against noise, as evidenced by lower Equal Error Rates. This suggests that integrating comprehensive noise profiles during the training phase is essential for developing more robust biometric systems.

The experimental findings also stressed the importance of the training method, with Triplet loss training showing superior performance in terms of prediction times and robustness compared to Softmax training. This insight is particularly relevant for applications requiring real-time processing.

The real-world testing scenarios further demonstrated that proximity to the microphone

MOST WIEDZY Downloaded from mostwiedzy.pl

significantly affects model accuracy, indicating the need for adaptive systems that can maintain performance across various user interactions.

In conclusion, this study serves as a foundation for future research aimed at addressing the identified gaps in voice biometric technology. Further exploration into adaptive parameterization methods, sophisticated noise augmentation strategies, and the potential of advanced deep learning architectures is recommended to enhance the accuracy, efficiency, and robustness of voice biometric systems. This will not only improve the performance in controlled environments but also ensure reliability in real-world, acoustically dynamic settings.

Acknowledgments:

This research was co-financed by the Polish National Centre for Research and Development (NCBR) under the European Regional Development Fund, The Smart Growth Operational Pro-gramme, Priority axis I: Support for R&D activity of enterprises, Project No. POIR.01.01.01-00-0785/21.

The authors wish to thank the staff of the Ambient System sp. z o.o. for their support and for providing the equipment and the rooms necessary to perform the experiments.

References

1. Abdul, Z.K., Al-Talabani, A.K.: Mel Frequency Cepstral Coefficient and its Applications: A Review. IEEE Access. 10 (November), 122136–122158 (2022)
2. Al-Karawi, K.A.: Mitigate the reverberation effect on the speaker verification performance using different methods. Int. J. Speech Technol. 24 (1), 143–153 (2021)
3. Ayoub, B., Jamal, K., Arsalane, Z.: Gammatone frequency cepstral coefficients for speaker identification over VoIP networks. In: 2016 International Conference on Information Technology for Organizations Development (IT4OD). pp. 1–5. (2016)
4. Chauhan, N., Isshiki, T., Li, D.: Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database. In: 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS). pp. 130–133. (2019)
5. Chen, X., Zahorian, S.A.: Improving speaker verification in reverberant Environments. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. pp. 5854–5858 (2021)
6. Chougala, M., Kuntoji, S.: Novel text independent speaker recognition using LPC based formants. In: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). pp. 510–513. (2016)
7. Desplanques, B., Thienpondt, J., Demuynck, K.: ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. 2020-Octob 3830–3834 (2020)
8. Kumar, T., Bhukya, R.K.: Mel Spectrogram Based Automatic Speaker Verification Using GMM-UBM. In: 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). pp. 1–6. (2022)
9. Kumar, T., Bhukya, R.K.: Mel Spectrogram Based Automatic Speaker Verification Using GMM-UBM. 9th IEEE Uttar Pradesh Sect. Int. Conf. Electr. Electron. Comput. Eng. UPCON 2022. 1–6 (2022)
10. Leu, F.Y., Lin, G.L.: An MFCC-based speaker identification system. Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA. 1055–1062 (2017)
11. Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., Zhu, Z.: Deep Speaker: an End-to-End Neural Speaker Embedding System. ArXiv. abs/1705.0 (2017)
12. Ma, Y., Lee, K.A., Hautamäki, V., Ge, M., Li, H.: Gradient Weighting for Speaker Verification in Extremely Low Signal-to-Noise Ratio. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 11311–11315 (2024)
13. Maazouzi, A., Aqili, N., Aamoud, A., Raji, M., Hammouch, A.: MFCC and similarity measurements for speaker identification systems. Proc. 2017 Int. Conf. Electr. Inf. Technol. ICEIT 2017. 2018-Janua 1–4 (2018)
14. Misra, S., Das, T., Saha, P., Baruah, U., Laskar, R.H.: Comparison of MFCC and LPCC for a fixed phrase speaker verification system, time complexity and failure analysis. IEEE Int. Conf. Circuit, Power Comput. Technol. ICCPCT 2015. 7–10 (2015)

MOST WIEDZY Downloaded from mostwiedzy.pl

15. Naveen, R., Jeevan Reddy, C., Tanguturu, R., Anand Kumar, M.: Speaker Identification and Verification using Deep Learning. In: Int. Conf. Signal Inf. Process, pp. 1–6 (2022)
16. Novoselov, S., Lavrentyeva, G., Avdeeva, A., Volokhov, V., Gusev, A.: Robust Speaker Recognition with Transformers Using wav2vec 2.0. ArXiv. abs/2203.1 (2022)
17. Oglic, D., Cvetkovic, Z., Sollich, P.: Learning Waveform-Based Acoustic Models Using Deep Variational Convolutional Neural Networks. IEEE/ACM Trans. Audio Speech Lang. Process. 29 2850–2863 (2021)
18. Oglic, D., Cvetkovic, Z., Sollich, P., Renals, S., Yu, B.: Towards Robust Waveform-Based Acoustic Models. IEEE/ACM Trans. Audio Speech Lang. Process. 30 1977–1992 (2022)
19. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Libripeech: An ASR corpus based on public domain audio books. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 2015-Augus 5206–5210 (2015)
20. Qin, X., Bu, H., Li, M.: HI-MIA: A Far-Field Text-Dependent Speaker Verification Database and the Baselines. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7609–7613. (2020)
21. Ravi, V., Fan, R., Afshan, A., Lu, H., Alwan, A.: Exploring the use of an unsupervised autoregressive model as a shared encoder for text-dependent speaker verification. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. 2020-Octob 766–770 (2020)
22. Recasens, D.: A cross-language acoustic study of initial and final allophones of /l/. Speech Commun. 54 (3), 368–383 (2012)
23. Rémy, P.: Deep Speaker: An End-to-End Neural Speaker Embedding System - Unofficial Tensorflow/Keras implementation of Deep Speaker, https://github.com/philipperemy/deep-speaker, Accessed: , (2020)
24. Rusli, A.T., Ahmad, M.I., Ilyas, M.Z.: Improving speaker verification using MFCC order. Proc. 2016 Int. Conf. Robot. Autom. Sci. ICORAS 2016. 1–4 (2017)
25. Saritha, B., Laskar, M.A., K, A.M., Laskar, R.H., Choudhury, M.: CACRN-Net: A 3D log Mel spectrogram based channel attention convolutional recurrent neural network for few-shot speaker identification. Comput. Electr. Eng. 115 109100 (2024)
26. Shi, X., Yang, H., Zhou, P.: Robust speaker recognition based on improved GFCC. In: IEEE International Conference on Computer and Communications, pp. 1927–1931 (2016)
27. Shon, S., Tang, H., Glass, J.: Voiceid loss: Speech enhancement for speaker verification. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, pp. 2888–2892 (2019)
28. Singh, M.K.: A text independent speaker identification system using ANN, RNN, and CNN classification technique. Multimed. Tools Appl. (0123456789), (2023)
29. Song, S., Zhang, S., Schuller, B.W., Shen, L., Valstar, M.: Noise Invariant Frame Selection: A Simple Method to Address the Background Noise Problem for Text-independent Speaker Verification. Proc. Int. Jt. Conf. Neural Networks. 2018-July (2018)
30. Tan, C.K., Irving, P.E., Mba, D.: A comparative experimental study on the diagnostic and prognostic capabilities\nof acoustics emission, vibration and spectrometric oil analysis for spur gears. 21 (1), 208–233 (2007)
31. Tazi, E.B.: A robust Speaker Identification System based on the combination of GFCC and MFCC methods. In: 2016 5th International Conference on Multimedia Computing and Systems (ICMCS). pp. 54–58. (2016)
32. Traer, J., McDermott, J.H.: Statistics of natural reverberation enable perceptual separation of sound and space. Proc. Natl. Acad. Sci. U. S. A. 113 (48), E7856–E7865 (2016)
33. Valero, X., Alias, F.: Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification. Trans. Multi. 14 (6), 1684–1689 (2012)
34. Yu, Y.Q., Li, W.J.: Densely connected time delay neural network for speaker verification. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. pp. 921–925 (2020)
35. Zhao, Z., Duan, H., Min, G., Wu, Y., Huang, Z., Zhuang, X., Xi, H., Fu, M.: A lighten CNN-LSTM model for speaker verification on embedded devices. Futur. Gener. Comput. Syst. 100 751–758 (2019)
36. Zhao, Z., Li, Z., Wang, W., Zhang, P.: PCF: ECAPA-TDNN with Progressive Channel Fusion for Speaker Verification. In: IEEE Int. Conf. Acoust. Speech Signal Process. pp. 1–5 (2023)

MOST WIEDZY Downloaded from mostwiedzy.pl