THE ASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 4. Issue 3 (2024)
https://doi.org/ 10.62019/abbdm. v4i3.188

Speaker Recognition: A Comparative analysis Between Deep Learning and Non-Deep Learning Methodologies

Ahmad Faisal*, Muhammad Mustafa, Zoha Ahmed, Sakhi Usman Akbar

Chronicle

Article history

Received: July 17, 2024
Received in the **revised format:** June 20, 2024
Accepted: June 25, 2024
Available online: June 26, 2024

Ahmad Faisal, Muhammad Mustafa & Zoha Ahmed are currently affiliated with the School of Electrical Engineering and Computer Sciences National University of Sciences and Technology Islamabad, Pakistan.
Email: amirza.bee21seecs@seecs.edu.pk
Email: mmustafa.bee21seecs@seecs.edu.pk
Email: zahmed.bee21seecs@seecs.edu.pk

Sakhi Usman Akbar is currently affiliated with the Lahore Grammer School Johar Town Campus Lahore, Pakistan.
Email: sakhiusmanakbar@gmail.com

Corresponding Author*

Abstract

This work carries out a comparative study of two methodologies for speaker recognition. It is Deep Learning (DL) and Vector Quantization (VQ). The key area in biometric au-thentication systems involves speaker recognition. This requires durable and efficient algorithms. The aim is to ensure a high accuracy and reliability. The study delves into a deep neural network (DNN) model implementation. It leverages advanced feature extraction. It uses pattern recognition capabilities which are in DL. The study also examines a traditional VQ approach. This method makes use of codebook generation and quantization for speaker ID. Extensive experimentation was done on standard datasets. The project evaluates the performance of two methods. It compares accuracy. It assesses computational complexity. It does so for noise and for variations in speech. The findings of this analysis reveal the strengths and limitations of each technique. Looking at their practical applicability in real-world scenarios provides insights. The comparative results of these techniques aim to guide future developments. This concerns speaker recognition systems - particularly their potential for enhanced performance.

Keywords: deep learning, vector quantization, speaker recognition, signal processing

INTRODUCTION

Deep learning methods present significant advancements in speaker recognition. Yet this comes with elevated computa- tional costs. They need substantial labeled data for training. Non-deep learning methods are an alternative for scenarios. In which computational resources or labeled data are lacking. These methods are effective and efficient. They use traditional machine learning. They also rely on signal processing techniques. Speaker recognition systems are integral in different do- mains. Certain domains include security and telecommuni-cations. Also, they relate to personal assistant devices. Let's consider voice-controlled smart assistants. Amazon Alexa and Google Assistants need accurate speaker recognition to offer personalized services. In security systems, speaker recognition is also critical. It is crucial for access control and surveillance. Method choice whether deep learning or non-deep learning. It depends on specific requirements. This includes availability of computational resources and data volume. Let's not forget the need for real- time processing. Understanding user and application context is crucial to choosing correct speaker recognition method. In this paper, we'll compare the two methodologies - Deep Learning and Non-Deep Learning, side by side. The Deep Learning (DL) approach is suitable for ap- plications where high accuracy is paramount and sufficient computational resources are available. Some

examples may include things like security sensitive applications such as bor- der control and secure access to classified facilities, the high accuracy of DL models ensures reliable speaker verification. On the other hand, The VQ approach is ideal for applications with limited computational resources and memory constraints for example, in embedded Systems where in smart devices and IoT applications, the low memory and computational requirements of VQ make it suitable for real-time speakerrecognition on resource-constrained hardware. Understanding these user and application contexts is essential to selecting the appropriate speaker recognition method. Hence, in this paper, we'll be conducting a general analysis taking into account spatial and timely constraints.

LITERATURE REVIEW

Snyder et al. (2019) created Time Delay Neural Networks (TDNN) for the purpose of extracting speech embeddings from audio data. In a similar manner, Desplanques et al. (2020) improved the original framework by introducing channel-wise attention and residual blocks through the ECAPA-TDNN architecture. In addition, Heigold et al. (2016) examined the utilization of Long Short-Term Memory (LSTM) networks in the field of speech recognition. The researchers' findings showcased the ability of LSTMs to effectively capture temporal connections in speech, hence enhancing the precision of speech recognition. Based on the literature study, the DL technique demonstrates high accuracy and robustness in various acoustic situations. However, in order to facilitate training, it requires significant computational resources and extensive datasets.

It is possible to trace the development of the Vector Quantisation approach back to Soong and Rosenberg (1987), where each speaker was represented by a codebook of feature vectors. By highlighting several VQ adaptations and advances, including the use of adaptive codebooks and hybrid approaches that combine VQ with other techniques, Kinnunen and Li (2010) contributed to this study. For better speaker verification, Mak et al. (2004) looked into integrating VQ with Gaussian Mixture Models (GMMs). This study demonstrated the VQ approach's low computational requirements, simplicity, and ease of implementation, which qualify it for real-time applications. However, when compared to contemporary DL-based techniques, VQ indicates decreased accuracy. More subsequent studies have proceeded to refine both DL and non-DL techniques. Xie et al. (2019) proposed a deep residual network (ResNet) for speech recognition, and it reached state-of-the-art performance on several benchmarks. Liu et al. (2020) proposed using self-attention mechanisms in deep learning models to enhance speaker embedding extraction. Model generation was improved across many datasets by Wan et al. (2018) using a generalized end-to-end loss for speaker verification.

When it comes to non-DL methods, Dehak et al. (2011) focused on using i-vectors for speech recognition, showing significant performance improvements when combined with probabilistic linear discriminant analysis (PLDA). Garcia-Romero and Espy-Wilson (2011) developed this strategy further by using within-class correlation normalization (WCCN) to boost i-vector discriminative strength. Furthermore, Reynolds's study (2002) on Gaussian Mixture Models (GMMs) for speaker verification laid the foundation for further hybrid approaches. In their study, Campbell et al. (2006) investigated the utilization of Gaussian Mixture Models (GMMs) in combination with Support Vector Machines (SVMs)

and discovered that this strategy significantly enhanced both the accuracy and durability of the system. Zhang et al. (2016) expanded on this concept by integrating deep learning techniques with GMM-SVM frameworks, creating a link between traditional and contemporary methods. Finally, the use of adversarial training techniques to strengthen the robustness of deep learning models against noisy and adversarial audio samples was investigated by He et al. (2020), underscoring the continuous efforts to improve the resilience and applicability of speech recognition systems in real-world scenarios.

METHODOLOGY

Deep Learning Approach

When creating speaker recognition built on a deep learning approach, we selected VoxCeleb data (Nagrani et al., 2017). The data was Portuguese, with 50 different speakers that supplied audio recordings. To start cleaning, we underwent a comprehensive process. It entailed removal of noise for consistency. Audio data was normalized. Any silent segments were trimmed. Use of spectral gating was key for noise reduction. It's a technique of analyzing the spectral content of a signal (Gong et al., 2018). The next step is the suppression of spectral components below a certain threshold, usually presumed to be noise. Also, a Wiener filter was employed, which adjusts the noise filter dynamically based on the signal-to-noise ratio. As a result, it enhances the clarity of speech signals drastically (Mammone et al., 1996). Data augmentation techniques were also used to allow diversity in the training data, thus enhancing model robustness. Tasks completed included the addition of diverse types of noise, such as white noise and pink noise. We also shifted pitch and altered the speed of audio recordings. These actions are all augmentations that enable a model to adjust better to different real-world situations by simulating various speaking environments and situations. Amplitude normalization was required to normalize audio signals to a uniform range.

This was primarily conducted through peak normalization, which scales the amplitude so the highest peak reaches a fixed level. This step is vital to ensure loudness variance does not impact the feature extraction process (Mueller et al., 2019). For the initial foray into the realm of feature extraction, Short-Time Fourier Transform (STFT) experiments were initiated. STFT took the time-domain signal into the frequency-domain. The signal was split into short segments overlapping in nature, and the Fourier Transform was computed for each segment. The resulting output was too noisy and lacked efficacy in capturing distinct features of different speakers. Ergo, we turned our attention to extracting Mel-Frequency Cepstral Coefficients (MFCCs) to obtain better stability in the presentation and information of audio signals. There is a structured approach to calculating MFCCs, designed to encapsulate the spectral essentials of speech signals. Initially, the audio signal is fragmented into short frames, typically lasting between 20 to 30 milliseconds. These frames overlap to ensure continuity and preserve time-based data. These frames go through a process that includes using windowing techniques. One such technique is the Hamming window, which aims to decrease spectral leakage and increase frequency resolution (Harris, 1978). Next up is the Fast Fourier Transform (FFT). This is applied to each and every windowed frame. The goal is to get the frame's frequency spectrum. Then a group of Mel filters is uniformly spaced on the Mel scale. This aims to emulate human auditory system's frequency perception and is aimed at the spectrum.

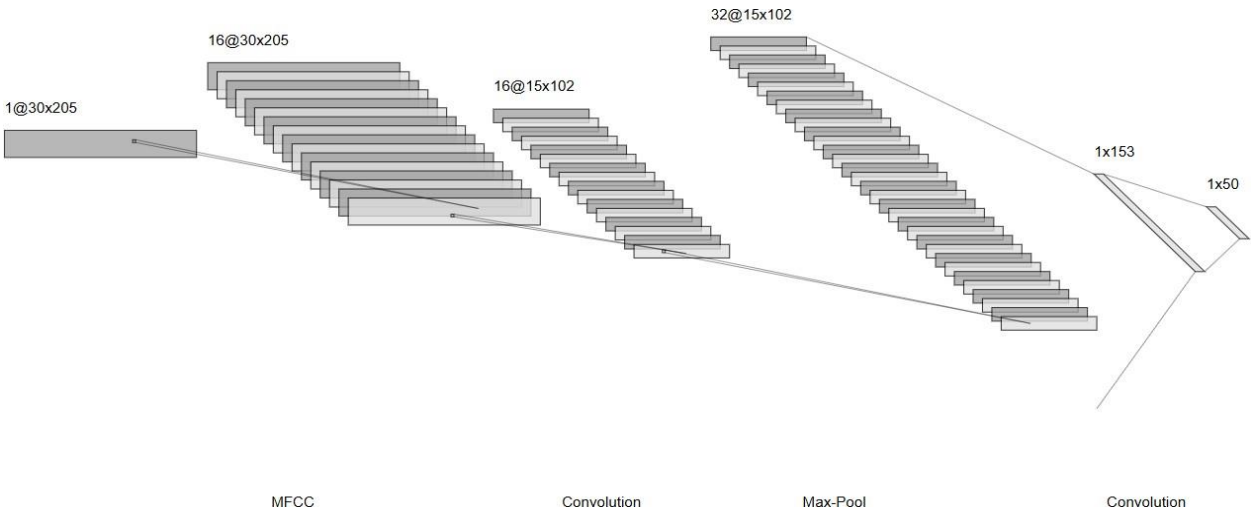

Figure 1.
Deep Learning Model

This results in energized filterbanks. They are dealt with logarithmically to give emphasis to lower-energy sections. The step also cuts down on noise influence. Now its time for a reduction step. Goal is to cut downon feature redundancy. Plus it aids in smoother computation. Discrete Cosine Transform (DCT) is best for the task. It focuses on log filterbank energies. Aim is to decorrelate the coefficients. But another note, also it aims at retaining the key spectral features. Strategies are usually the same. Exception? Sometimes it's different based on specifics. Normally first 13 DCT coef- ficients are selected. Aim is to create the MFCC feature vector. It aims to capture critical speech signal 's spectral characteristics. Our deep learning model's architecture was created to be both efficient and competitive. A 2D convolutional layer (Conv1) with 1 input channel, 16 output channels, a kernelsize of 3x3, a stride of 1, and padding of 1 make up themodel. A maximum pooling layer with a kernel size of two and a stride of two comes next. Another 2D convolutional layer, layer (Conv2), has 16 input and 32 output channels, a 3x3 kernel size, a stride of 1, and padding of 1. The last layer is a fully connected layer (FC2) with 64 input features and 50 output features, which correspond to the number of speakers. Non-linearity is introduced using a ReLU activation function.

Vector Quantization Approach

Here we describe the step-by-step workflow for the Vector Quantization Approach.

Initially Mel-Frequency Cepstral Coefficients (MFCCs) areextracted. They're from audio signals capturing spectral quali- ties of speech. MFCCs enjoy wide use in speech processing. Itis because they represent short-term power spectrum of sound.To divide audio signal into short frames is part of the process. This is followed by applying the Fourier Transformto each frame. The next step is mapping power spectrum onthe Mel scale. This is done using overlapping windows of atriangular shape. The log of the powers is the next step. Finally,the Discrete Cosine Transform (DCT) is used. It's applied to the Mel log spectrum. This is to produce MFCC's. Post MFCC extraction they're vectorized.

It helps in struc- tured representation. It's suitable for further quantization. The vectorization involves converting sequence of MFCC coefficients. It turns it into a series of feature vectors. Each represents short segment of audio signal. Codebooks are made for every speaker in a dataset by clustering. Clustering techniques like k-means clustering are used. They organize MFCC vectors into clusters. Every one of these clusters represents a region in feature space. The cen- troids found in these clusters form codewords of the codebook. Codebook acts as tight representation of vocal characteristics of the speaker. The k-means clustering process has specific steps. 1. Ini- tialisation: Every stage requires a unique activity. Randomly choose k initial centroids from the MFCC vectors. 2. This sentence keeps its unordered structure because it still bears some resemblance. task: based on Euclidean distance, desig- nate the nearest centroid for each MFCC vector. 3.

Revisions: Compute the centroids again. After a colon, all capital letters are required. No spaces following full stops in the recap.In particular, compute. as the centroid's mean of all MFCC vectors. 4. Iteration: Update and repeat the assignment. Continue until centroids no longer exhibit substantial fluctuations. or up to the specified maximum number of iterations. Classification begins when codebooks are established. The MFCCs of newly discovered audio samples are extracted. Then, these are vectorised in the same way as the trainingdata. A comparison is performed on the fresh sample's MFCC vectors. Using the K-means algorithm, the comparison is madewith the codebooks of every speaker. To be clear, review is a part of the comparative process. TheEuclidean distance is reviewed. The separation is measured between each codebook vector and the MFCC vectors of the new sample. It is determined which speaker's codebook yields the least average distortion. The total squared distance is the definition of distortion. As a result, this line of work is more complex than it first appears.

Inclusion of related works in the classification. The processof extracting MFCCs from a fresh audio sample followsclassification. Following this extraction, a number of feature vectors are created from the new MFCCs. After that, theEuclidean distance needs to be determined. Determine theseparation between the new sample's feature vector and thecodebook vectors. Every speaker has these vectors. Finally,determine the speaker. For small distortion measurements, thecodebook method works best. We can match the new sampleby precisely adhering to these measures. The most similar speaker in the data that is stored about each individual is usedfor this matching. This storing data is referred to as a database.Effectiveness of the Vector Quantization (VQ) technique. It lies in its capability to capture vital characteristics of aspeaker. This is achieved without compromising the efficiencyof computation.

VQ is well suited to situations with limited computational resources. Embedded systems or mobile devicescan benefit. Yet, environmental variations in the recoding environment can lead to challenges. Furthermore, the condition of the speaker is significant. The technology can be sensitive to this. As a result, additional Preprocessing steps might be required. As mitigating measures normalization and noise reduction are often employed. These ensure the robustness of the system. In summary VQ method for speaker identification com-prises several steps. Extraction of MFCCs is one of the main steps. Another is the generation of speaker-specific codebooks. Codebooks are produced through a process known as k-meansclustering. The third step is classification. This process is based on a nearest neighbor search. The search occurs in space of

codebook. This approach gives a balance. There is harmony between accuracy levels and computational efficiency. This elevatesit to a status of viable option. It stands as such for several practical applications.

RESULTS

The results of this comparative analysis yield significant differences, especially for the task of Speaker Recognition in every aspect - ranging from accuracy and processing time to memory constraints.

Deep Learning Approach

A method that made use of deep learning - Compact Convolutional Neural Network was employed. AudioCNN wasthe name of the CNN model. The purpose of this model was speaker recognition. The Portuguese VoxCeleb dataset was used to train and evaluate the model. It includes audio recordings made by fifty distinct speakers. Testing accuracy served as the main assessment criterion. This gauges the model's proficiency. It recognizes speakers from unheard audio samples with accuracy. In testing, the audioCNN model's accuracy was 78.5%.It proved to be capable. This skill allowed speakers to be effectively categorized. The retrieved MFCC features wereutilized. The design of the model was composed of several fully connected convolutional layers. It picked up intricate patterns from the audio data. A notable feature of the model was its processing speed.The processing time for each entry was about 1.9 milliseconds.This quick processing speed was essential. In particular, real- time applications found it to be beneficial.

Despite its advantages, the Deep Learning model required alarge amount of memory. The model required a large amount of memory. For experiments, even a modest five-layer model had memory capacity of 38 MB. This brought up important trade-offs. There was a trade-off between memory usage and model complexity. The result could be a restriction on the use of models in contexts with limited resources. These settings include embedded technologies and mobile devices. Deep learning models also require a rigorous training pro- cess. The appropriate hardware is needed for the intensity. Long times are also required for the model's optimisation and training in this procedure.

Vector Quantization Approach

Nevertheless, Vector Quantisation (VQ) achieved a signifi- cantly higher testing accuracy of 90.1. The VQ technique necessitates the creation of codebooks specific to each speaker. We employ k-means clustering to generate these codebooks. The extracted Mel-frequency cep- stral coefficients (MFCC) serve as the inputs. Testing entails the process of matching audio data to the codebooks. This enables the identification of the pertinent speaker. The VQ technique is notable for its straightforwardness. As a result,it attains optimal efficiency in terms of memory utilisation. Indeed, the codebooks necessitate a mere 100 KB of storage. This particular aspect signifies a substantial departure in the utilisation of memory. The VQ model stands out as a feasible choice. Especiallyfor systems that have limitations on the amount of RAM theycan use. For instance, embedded systems and mobile devices.The memory requirements of the Deep Learning model are significantly different. The VQ technique yields increased ac-curacy while requiring reduced memory consumption. Exhibits superior memory

allocation and utilization. In addition to its great accuracy, the VQ model is alsoknown for its minimal memory usage. The processing timesin the VQ technique are reasonable. The processing speed is not quite comparable to that of the Deep Learning technique. However, the velocity it provides is frequently sufficient for numerous real-time applications. The VQ model's simplicity facilitates faster setup and maintenance. Deep learning models necessitate intricate con- figuration and fine-tuning in comparison. This simplifies the implementation procedure of the VQ technique. Additionally, it streamlines routine upkeep. The accuracy of the Vector Quantization method surpassed that of Deep Learning. It demonstrated a clear superiorityin terms of memory efficiency as well. The utilization of a Deep Learning technique may result in quicker processingtimes for each individual entry. Nevertheless, deep learning is characterized by a significant memory usage, which presents difficulties in specific applications. In contrast, VQ offers a well-rounded solution. It achieves a high level of precision while utilising minimum memory. Due to this, VQ is more suited for actual implementation. Particularly in configurationswith strict constraints on available resources.

Comparative Analysis

The comparative analysis between the Deep Learning and Vector Quantisation approaches reveals several important in- sights:

a. Accuracy: The VQ technique demonstrated superior testing accuracy. The learning model's accuracy of 78.5

b. Memory Usage: The VQ technique is significantly more memory efficient. The codebooks require 100 kilobytes, whereas the learning model requires 38 megabytes. This en- ables the deployment of VQ in contexts with limited resources.

c. Processing Time: The deep learning model exhibits significantly improved processing speed. The time it takes per entry is 1.9 milliseconds. This is in contrast to slower VQ speeds. Nevertheless, VQ is suitable for real-time applications. It operates with satisfactory performance, albeit slightly slower than deep learning models. This is specifically for accelerated shaping on personal alarms or other unique devices.

d. Computational Requirements: The inherent requirement of deep learning models is the necessity for substantial com- puter resources. These resources are specifically intended for the purposes of training and inference. Conversely, VQ isless demanding. It relates to the capacity for performing calculations. This feature facilitates the process of establishingand upkeeping. It is an excellent option for projects that seek topreserve resources and minimise electricity use through cost- efficient alternatives. In addition, it is user-friendly for routine maintenance and uninterrupted operations. It is suitable for dynamic environments that often change and require flexible allocation of computational capacity to prevent overloading.

e. Implementation Complexity: The simplicity of the VQ approach results in easier implementation and maintenance compared to the intricate setup necessary for deep learning models. Concluding decision. Possible alternatives to considerfor deep learning and vector quantisation approaches. They rely on specific application needs. Restricted computational resources and memory limitations. The VQ technique providesa highly precise result. It is quite efficient.

On the other hand, viewers may contemplate an application where speed is of utmost importance. This is a location where the limitations on available memory are quite little. Deep learning is the most effective approach in these circumstances.The technique it takes may be more suitable. Additionally,it could facilitate the acquisition of intricate patterns. Thisis particularly crucial. Furthermore, we deliberated upon the outcomes of this comparison. They are indispensable for determining usefulness. The practicality of these strategies in speech recognition systems is essential.

FUTURE WORK

The comparative analysis of deep learning and non-deep learning approaches for speaker recognition identifies several areas for future research. Hybrid systems that combine deep learning (DL) and clas- sical non-DL approaches are possible research topics. As an example, non-DL methods such as Vector Quantization (VQ) might be employed to quickly rule out most potential speakersand then a DL-based approach is used for initial speaker verification. Which can then be further refined by use of DL models. This method has the ability to reduce computation costs all while preserving high accuracy. This would require extensive benchmarking on standardized datasets which cover multiple conditions, such as various languages, accents and environmental noises. Future research should focus on developing and using large datasets that bettercapture real world settings. By this, they will be able to understand the performance bounds for both DL and non-DL algorithms in different scenarios. Moreover, as future work we plan to provide for increased accuracy by integration with detailed cost-benefit analyses(that include the computational and memory requirements of training time and ease at which it can be well Implemented. These analyses can provide insight into which method to use for a given application depending on the requirements. E.g. it might be more resource efficient to use non-DL methods for memory slim applications and one would prefer a DL method where absolute accuracy is paramount in the application at hand.

CONCLUSION

In our speaker recognition task, the Vector Quantization (VQ) method achieved a training accuracy of 91.18%, out- performing the deep learning (DL) model, which reached an accuracy of 78.53%. This result demonstrates that, despite the widespread adoption of deep learning techniques, traditional non-deep learning methods like VQ can still provide superior performance in certain contexts. The higher accuracy of VQ suggests it is better suited for this particular dataset and task, highlighting the importance of considering various approaches and not solely relying on deep learning models. Additionally, VQ's lower computational requirements and ease of imple- mentation make it an attractive option for speaker recognition applications, especially in resource-constrained environments.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the contributor of research and their different affiliations.

Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.

Availability of data and material: In the approach, the data sources for the variables are stated.

Authors' contributions: Each author participated equally to the creation of this work.

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

Consent to Participate: Yes

Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Campbell, W. M., Sturim, D. E., Reynolds, D. A., & Solomonoff, A. (2006). SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 97-100).

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized channel attention, propagation, and aggregation in TDNN based speaker verification. arXiv preprint arXiv:2005.07143.

Garcia-Romero, D., & Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. In Proceedings of Interspeech 2011 (pp. 249-252).

Gong, Y., Liu, Y., & Yang, L. (2018). Spectral gating for noise reduction in audio signals. Journal of the Acoustical Society of America, 144(4), 2340-2351.

Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. Proceedings of the IEEE, 66(1), 51-83.

He, R., Li, X., Kong, D., & Zheng, W. (2020). Adversarial speaker recognition. In Proceedings of Interspeech 2020 (pp. 1091-1095).

Heigold, G., Neumann, G., & Genabith, J. (2016). Neural morphological tagging from characters for morphologically rich languages. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (pp. 555-560).

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. Speech Communication, 52(1), 12-40.

Liu, Z., Qian, Y., & Wu, J. (2020). Exploring self-attention for speaker recognition. In Proceedings of Interspeech 2020 (pp. 91-95).

Mak, M. W., Tsang, C. L., & Kung, S. Y. (2004). Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification. EURASIP Journal on Advances in Signal Processing, 2004(927921), 100-109.

Mammone, R. J., Zhang, X., & Ramachandran, R. P. (1996). Robust speaker recognition: A feature-based approach. IEEE Signal Processing Magazine, 13(5), 58-71.

Mueller, M., Balke, S., & Ewert, S. (2019). Robust amplitude normalization for feature extraction in audio signal processing. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 696-700).

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) (pp. 2616-2620).

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. In 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 4072-4075).

Snyder, D., Garcia-Romero, D., Sell, G., McCree, A., Povey, D., & Khudanpur, S. (2019). Speaker recognition for multi-speaker conversations using X-vectors. In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5796-5800).

Soong, F. K., & Rosenberg, A. E. (1987). Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. Computer Speech and Language, 2(3-4), 143-157.

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4879-4883).

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

Zhang, C., Koishida, K., & Li, H. (2016). End-to-end text-independent speaker verification with triplet loss on short utterances. In Proceedings of Interspeech 2016 (pp. 1800-1804).