Speaker Identification

Audio and Speech Processing

Dino Dervisevic
University of Applied Sciences Munich

Author Note

The purpose of this paper is to provide a state-of-the-art report on the topic of speaker identification. To achieve this, I, the author Dino Dervisevic, have relied on 19 academic papers listed in the reference section. These papers form the foundation for all sections of this report, specifically from **Section I: Introduction** to **Section VI: Challenges and Solutions**. As such, the knowledge, ideas, and wording in these sections are derived from these referenced works, and I do not claim ownership of them. If one or more references are provided at the end of a paragraph or a larger text section, it indicates that the content has been paraphrased or synthesized from the information found in those references.

The only sections where I, Dino Dervisevic, have contributed my own original ideas are **Abstract**, **Discussing The Results Conclusion**, and **Future Directions**. In particular, the **Future Directions** section represents my impressions and insights, developed while learning, working on this paper and my professional background in generative artificial intelligence.

Abstract

Speaker identification is a biometric technology that identifies individuals based on their unique voice characteristics. This paper provides an overview of the field, outlining fundamental concepts, system architectures, and practical applications. It examines key components of speaker identification systems, including feature extraction techniques such as MFCC, FFT, and DWT, and modeling approaches like GMM, CNN, and LSTM. The paper also addresses significant challenges faced by these systems, such as variability in voice characteristics and background noise. Proposed solutions, including Multi-task Neural Networks (MTNs) and Noise Invariant Frame Selection (NIFS), are explored to enhance system robustness and accuracy. By analyzing existing techniques and their limitations, this work highlights ongoing advancements and promising research directions for building reliable and efficient speaker identification systems.

I. Introduction

Speaker identification, a subset of speaker recognition, involves determining an individual's identity based solely on their voice characteristics. [1]

The process of speaker identification typically encompasses several stages: feature extraction, modeling, and classification. Feature extraction involves capturing distinctive attributes from speech signals, which effectively represent the power spectrum of sound. These features are instrumental in distinguishing between different speakers. [2] Subsequently, speaker modeling employs techniques to create statistical representations of a speaker's vocal traits. The final classification stage matches input speech against these models to ascertain the speaker's identity. [1]

In recent years, deep learning has significantly advanced speaker identification systems. Deep neural networks (DNNs) excel at learning complex patterns in speech data, enabling breakthroughs in feature extraction and modeling. These advancements have broadened the scope of speaker identification to include challenging real-world scenarios, such as short-duration utterances, cross-lingual data, and noisy environments, emphasizing the transformative potential of deep learning in this domain. [2]

The applications of speaker identification are diverse. **Telephony Authentication for Transactions:** This technology secures automated phone services by verifying users' identities during transactions, enhancing security for banks and telecom companies. It's often combined with other methods to improve performance and prevent fraud, such as requiring users to repeat system-generated phrases to thwart pre-recorded attacks. Additionally, it's used in monitoring programs for home incarceration and parole. **Access Control:** Integrating voice verification with traditional security measures like keys or badges enhances physical access control. Applications include voice-activated door locks and vehicle ignition systems, as well as secure access to computers and mobile devices. These systems typically use specific passwords for verification. **Speech Data Management and Personalization:** Identifying who is speaking and when helps organize audio content, benefiting media industries through automatic speaker indexing and subtitling. It's also useful for annotating meeting recordings and personalizing services based on characteristics like gender or age. **Forensic Speaker Recognition:** In criminal investigations, analyzing voice recordings can assist in identifying individuals. However, the scientific community agrees that such analyses should not be the sole basis for determining guilt or innocence. These applications highlight the versatility of speaker identification in enhancing security, personalizing user experiences. [1]

Speaker identification faces several challenges. There are three main reasons for this. First, a person's voice can change a lot depending on their mood, health, or age. Second, the differences between people's voices, especially among family members, are not always strong enough to make them easy to tell apart. Finally, background noise and issues with recording equipment can mess up the voice data, making it less reliable. To deal with these problems, people sometimes need to record their voices multiple times or for longer periods, but this can be inconvenient. [1] [14]

In summary, speaker identification is a vital technology with significant implications for security, forensics, and human-computer interaction. Continuous research and technological progress are essential to address existing challenges and fully realize its potential.

II. General Overview

To understand speaker identification, it's essential to grasp the following key concepts.

1) **Terminology: Speaker Recognition** is the overarching process of identifying or confirming a person's identity using their voice characteristics. This field is divided into two primary tasks: speaker identification and speaker verification. **Speaker Identification** involves determining which individual, from a set of known speakers, matches a given voice sample. This process answers the question, "Who is speaking?" It's a one-to-many comparison, where the system compares the input voice against multiple stored voice prints to find the best match. In contrast, **Speaker Verification** is the process of confirming whether a speaker's claimed identity is genuine. This answers the question, "Is the speaker who they claim to be?" It's a one-to-one comparison, matching the input voice to a specific voice print associated with the claimed identity. [1] [3]
2) **Different Systems:** In speaker identification, systems are categorized as either **closed-set** or **open-set**, based on the inclusion of potential speakers in the system's database. In a **closed-set** scenario, the system operates under the assumption that the speaker is among the known, enrolled individuals. The task involves matching the input voice sample to one of these known speakers, effectively performing a one-to-many classification. Conversely, **open-set** speaker identification acknowledges the possibility that the speaker may not be in the existing database. The system must determine whether the input voice belongs to a known speaker or an unknown individual, labeling it as "unknown" if no sufficient match is found. [1] [4]
3) Speaker identification systems are categorized into two main types: text-dependent and text-independent. In **text-dependent systems**, the system is aware of the specific text that the speaker will say, and the speaker is required to cooperate by speaking this text. These systems can achieve high accuracy even with short speech samples because of the consistency in the spoken content. [3] [11] [12] On the other hand, **text-independent systems** do not rely on any prior knowledge of the spoken text. The system must work with any spoken content. However, these systems typically require longer speech samples to build reliable models and achieve good performance. [3] [12] [16]

By understanding these technical foundations, one can appreciate the complexities involved in speaker identification.

III. Data

Speaker identification systems are typically tested using datasets that include audio recordings of speech from multiple speakers, along with corresponding speaker labels. These datasets are essential for evaluating system performance under different conditions, such as varying background noise levels, accents, and speech durations. The recordings usually encompass controlled environments (e.g., studio recordings) and real-world scenarios (e.g., telephone calls or meetings). High-quality datasets ensure that the testing covers diverse speaker characteristics like age, gender, and linguistic backgrounds. [6] [7] [8] [9]

Several datasets are widely used for speaker identification research, each with unique features. **VoxCeleb1** and **VoxCeleb2** include thousands of speakers from YouTube videos, offering diverse, real-world speech data. These datasets are freely available upon request. [6] **Librispeech**, originally for speech recognition, contains high-quality audiobook recordings and is often reused for speaker identification. It can be downloaded for free from its website. [7] The **TIMIT** dataset includes 630 speakers from 8 U.S. dialect regions. Although small, its carefully labeled data makes it great for benchmarking. TIMIT is available for purchase through the LDC. [8] Lastly, the **CallFriend** dataset has telephone conversations in multiple languages, ideal for noisy, text-independent tasks. It also requires LDC membership or purchase. [9]

Once the recordings are gathered, they are annotated with speaker labels to associate each audio sample with its corresponding speaker identity. In some cases, additional metadata such as age, gender, and recording conditions is included to enable more detailed analysis and system training. [6] [7] [8] [9] To enhance robustness, synthetic data augmentation techniques, such as the addition of noise or reverberation, are frequently employed. These techniques help simulate challenging real-world environments, improving the generalizability of speaker identification systems. [10]

IV. State of the art

After collecting a suitable dataset of speech samples, speaker identification systems typically follow three main stages: Feature Extraction, Modeling, and Classification. These stages form the backbone of the identification process, transforming raw audio data into meaningful representations, building models to represent individual speakers, and ultimately determining the speaker's identity from new speech inputs. [3] [5] [13]

Feature Extraction: Feature extraction is a critical step in speaker identification systems, where the raw speech signal is analyzed to derive a compact, discriminative representation of the speaker's unique vocal traits. This process involves isolating speaker-specific characteristics from the speech signal while minimizing irrelevant information, such as background noise or spoken content. The extracted features capture fundamental properties of the voice, including physiological attributes like vocal tract shape and behavioral aspects such as speaking style. [5] [15] The goal of feature extraction is to reduce the dimensionality of the raw audio data while retaining essential information that distinguishes one speaker from another. This ensures that the subsequent stages, such as modeling and classification, can operate efficiently and accurately. [3] [14]

Modeling: The modeling stage in speaker identification systems involves creating a structured representation of a speaker's unique vocal characteristics. This representation is derived from the features extracted in the previous stage and serves as the basis for distinguishing between speakers. The purpose of modeling is to encapsulate the defining traits of each individual's voice in a way that the system can use for accurate identification. This stage is critical because it transforms raw feature data into a meaningful form that can be compared across different speakers. Speaker models also ensure scalability, enabling the system to identify a larger number of individuals while maintaining accuracy. [3] [5]

Classification: The primary goal of classification is to compute a similarity score or distance between the input features and the stored speaker models. Based on these scores, the system identifies the speaker whose model best matches the input data. This stage ensures the system's functionality in real-world scenarios, allowing it to handle variations in speech and environmental conditions effectively. [3] [13]

V. Evaluation of Feature Extraction and Modeling Techniques

In this section, we compare various feature extraction and modeling methods used for speaker identification systems. The goal is to assess their strengths, limitations, and trade-offs in terms of accuracy, computational efficiency, and suitability for handling different aspects of speech data. By exploring both feature extraction and classification techniques, this paper aims to provide an overview of the key factors essential for speaker identification systems.

Feature Extraction. The paper titled *"Comparison of Feature Extraction for Speaker Identification System"* [17] evaluates three widely used feature extraction methods—Fast Fourier Transform (FFT), Mel-Frequency Cepstral Coefficient (MFCC), and Discrete Wavelet Transform (DWT)—in the context of speaker identification systems. The study utilizes Dynamic Time Warping (DTW) as the classification method, allowing a comprehensive comparison of the three approaches.

1) **Mel-Frequency Cepstral Coefficients (MFCC)**: This method is inspired by the human auditory system, employing a non-linear frequency scale that mirrors how humans perceive sound. MFCC applies a linear scale for frequencies below 1 kHz and a logarithmic scale for frequencies above 1 kHz, effectively representing speech characteristics essential for speaker identification. The mathematical relationship between frequency (in Hz) and its corresponding Mel-frequency is expressed as:

$$\mathrm{Mel}(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right)$$

MFCC is widely recognized for its robust performance in speech and speaker recognition systems.

2) **Fast Fourier Transform (FFT)**: A computationally efficient algorithm used to transform signals from the time domain to the frequency domain. The FFT divides the signal into its even and odd components, using the Discrete Fourier Transform (DFT) equation:

$$X(k) = \sum_{n=0}^{\frac{N}{2}-1} x(2n) e^{-\frac{j2\pi(2n)k}{N}} + \sum_{n=0}^{\frac{N}{2}-1} x(2n+1) e^{-\frac{j2\pi(2n+1)k}{N}}$$

FFT is especially useful for analyzing stationary signals but may be less effective with non-stationary data.

3) **Discrete Wavelet Transform (DWT)**: Unlike FFT, DWT provides a dual-domain representation of signals by capturing both time and frequency characteristics. This is particularly advantageous for non-stationary signals like speech, which vary over time. DWT uses waveforms with limited duration and zero average value, defined as:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-\tau}{s}\right)$$

This flexibility makes DWT suitable for capturing signal anomalies and transient features.

The study involved recordings from 30 individuals (24 males and 6 females), with each participant uttering the same word twice, resulting in a dataset of 60 voice samples. FFT, MFCC, and DWT were applied to the recordings, followed by classification using DTW. The experimental results showed that DWT combined with DTW achieved the highest accuracy of 96.67%, outperforming MFCC and FFT, which both achieved 86.67% accuracy. This superior performance of DWT is attributed to its ability to analyze signals in both time and frequency domains simultaneously. However, DWT required a longer computational time compared to FFT and MFCC due to its larger feature vector size, demonstrating a trade-off between accuracy and efficiency.

Model Training. The paper titled *”Speaker Identification Using MFCC Feature Extraction: A Comparative Study Using GMM, CNN, RNN, KNN, and Random Forest Classifier”* [5] compares various modeling techniques for speaker identification. The comparison is based on the LibriSpeech dataset. The authors utilize Mel-Frequency Cepstral Coefficient (MFCC) as the common feature extraction method across all models, ensuring a consistent basis for comparison. The paper evaluates five speaker identification methods, which utilize MFCC features as input:

1) **Gaussian Mixture Model (GMM):** A statistical approach that models the distribution of MFCC features for each speaker using multiple Gaussian components, effectively capturing the unique speech characteristics.

2) **Convolutional Neural Networks (CNN):** A deep learning method that uses convolutional layers to learn local patterns from MFCC features, enabling robust speaker classification.
3) **Long Short-Term Memory (LSTM):** A type of Recurrent Neural Network (RNN) that captures temporal dependencies in sequential data, making it suitable for handling variable-length speech inputs.
4) **k-Nearest Neighbors (KNN):** A non-parametric algorithm that identifies the speaker by comparing MFCC features to the closest labeled neighbors based on feature similarity.
5) **Random Forest Classifier:** An ensemble method that combines multiple decision trees to classify speakers, offering good accuracy and interpretability while mitigating overfitting.

The results demonstrate that GMM achieves the highest accuracy at 98.68%, followed by LSTM at 95.77%, with KNN and Random Forest showing comparable performances. CNN performs slightly lower at 92%. The study highlights GMM's superior ability to capture statistical characteristics of speech, while deep learning models like LSTM effectively handle temporal dependencies in speech data.

VI. Discussing The Results

If the papers are interpreted as ground-truth, the following conclusions can be drawn:

1) **DWT:** Should be chosen when accuracy is the primary objective and computational resources are available, such as in offline or high-performance computing applications.
2) **MFCC and FFT:** Are better suited for real-time applications where computational efficiency is a key requirement.
3) **GMM:** Is ideal for tasks requiring high accuracy and statistical modeling of features, particularly in controlled environments.
4) **LSTM:** Is recommended for tasks involving sequential or temporal dependencies in speech, such as speaker identification in dynamic contexts.
5) **CNN, KNN, and Random Forest:** Are useful when simplicity, interpretability, or limited computational resources are priorities, albeit with some trade-offs in accuracy.

However the presented papers also showcase very grave weaknesses. In *"Comparison of Feature Extraction for Speaker Identification System"* [17] the results are based on a small dataset of 30 people (24 men, 6 women) with only 60 voice recordings. This is a problem because women are underrepresented, which can lead to bias. A model trained mostly on male voices might struggle to recognize female voices accurately. Future studies should use more balanced datasets with greater diversity in gender, age, and background. The feature extraction methods – FFT, MFCC, and DWT – each have their strengths and weaknesses. DWT performed the best because it can analyze both time and frequency aspects of speech. However, DWT is slow and requires more computational power, making it less practical for fast processing. FFT is faster but struggles with speech, which changes over time.

In *"Speaker Identification Using MFCC Feature Extraction: A Comparative Study Using GMM, CNN, RNN, KNN, and Random Forest Classifier"* [5] the GMM model achieved the highest accuracy (98.68%). It works well in controlled environments but might not perform as reliably in real-world scenarios, like noisy environments. Models like LSTM, which are better at handling time-based changes, could work better in such cases but need more computing power. Models like CNN and LSTM performed well because they were optimized for features like MFCCs, which was used in the feature extraction step. However, since the dataset consists only of audio books, it lacks diversity, which means these models might not perform as effectively in practical applications. This shows the risk of models being too tailored to the training data and not generalizing well.

Nonetheless, the papers provide valuable insights into understanding the trade-offs between different methods for speaker identification. They highlight the importance of selecting techniques that align with the specific requirements of the task, ensuring a balance between accuracy, complexity, and computational efficiency. However, they should not be used to define the best method or model.

VII. Challenges and Solutions

This section presents proposed solutions to the challenges faced by speaker identification systems. However, it is important to note that these solutions are not universal and are effective only within their specific areas of application.

1) **Variability in Voice Characteristics**: An individual's voice can fluctuate due to factors such as mood, health, or aging, complicating consistent identification. [1]
Proposed Solution: The paper *"Identity, Gender, Age, and Emotion Recognition from Speaker Voice with Multi-task Deep Networks for Cognitive Robotics"* [18] proposes Multi-task Neural Networks (MTNs) as a solution to challenges in speaker identification, particularly variability in voice characteristics caused by mood, health, or aging. MTNs utilize shared feature extraction backbones combined with task-specific branches to process related tasks, such as age estimation and emotion recognition, alongside speaker identification. This multi-task approach effectively models voice variability by leveraging contextual tasks to regularize and enhance the primary task of speaker identification. This method captures subtle and dynamic voice features while maintaining computational efficiency, making it suitable for real-time applications. [18]
2) **Background Noise**: Background noise significantly impacts the performance of speaker identification systems, often degrading accuracy due to mismatched acoustic conditions between training and testing environments. [1]

Proposed Solution: To address this challenge, the paper *"Noise Invariant Frame Selection: A Simple Method to Address the Background Noise Problem for Text-independent Speaker Verification"* [19] proposes the Noise Invariant Frame Selection (NIFS) method as a pre-processing technique to enhance system robustness. NIFS selects frames from speech data that are less affected by noise, ensuring that only high-quality, noise-invariant frames are used for model training and testing. Frames with the lowest distortion scores, indicating they are least impacted by noise, are ranked and selected as robust frames. These selected frames are then used to train and test speaker models, improving their performance under noisy conditions. The method is adaptable to different model architectures and features, including MFCCs and i-vector systems, making it highly versatile. [19]

VIII. Conclusion

Speaker identification systems have evolved significantly, leveraging advanced techniques in feature extraction, modeling, and classification. Despite these advancements, challenges remain, including variability in voice characteristics, familial vocal similarities, and environmental noise. The proposed solutions, such as MTNs for addressing voice variability and NIFS for enhancing robustness against noise, demonstrate the potential to improve system accuracy and reliability in real-world conditions. Feature extraction methods like DWT and classification models such as GMM and LSTM show strong performance in specific scenarios, but a careful balance between accuracy, computational cost, and system design is essential. Continuous innovation, combined with a focus on robust, adaptable solutions, will enable speaker identification systems to achieve broader applicability and reliability in diverse applications.

IX. Future Directions

Future work directions in speaker identification should focus on enhancing system robustness and scalability while addressing real-world challenges.

- **Integration with Multimodal Biometrics:** Combining voice with other biometric modalities, such as facial recognition or behavioral traits, provides a way to improve the precision and security of identification in complex scenarios while also mitigating the threat of deepfakes.
- **Low-Resource Environments:** Developing lightweight, computationally efficient models is critical for deploying speaker identification systems on edge devices and in resource-constrained environments.
- **Real-Time Applications:** Further optimization of real-time systems, particularly in dynamic environments such as smart homes and IoT devices, is essential.

By exploring these directions, the field of speaker identification will continue to advance, offering more reliable and versatile solutions for real-world applications.

References

[1] J. Hennebert, Li Stan Z. and Jain Anil "Speaker recognition, overview," in Encyclopedia of Biometrics, S. Z. Li and A. Jain, Eds. Boston, MA: Springer US, 2009, pp. 1262–1270, DOI: 10.1007/978-0-387-73003-5_198, Springer Link.

[2] R. Sharma, D. Govind, J. Mishra, A. K. Dubey, K. T. Deepak, and S. R. M. Prasanna, "Milestones in speaker recognition," *Artificial Intelligence Review*, vol. 57, no. 3, p. 58, Feb. 2024, DOI: 10.1007/s10462-023-10688 Springer Link.

[3] Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, Liisa Koivusalo, Sneha Das, Esteban Gómez Mellado, Mariem Bouafif Mansali, Daniel Ramos, Sudarsana Kadiri, Paavo Alku, and Mohammad Hassan Vali "Introduction to Speech Processing", 2nd Edition, 2022, DOI: 10.5281/zenodo.6821775, Online.

[4] M. Affek and M. S. Tatara, "Open-Set Speaker Identification Using Closed-Set Pretrained Embeddings," in Intelligent and Safe Computer Systems in Control and Diagnostics, Cham: Springer International Publishing, 2023, pp. 167–177, DOI: 10.1007/978-3-031-16159-9_14, Springer Link.

[5] D. R. Yerramreddy, J. Marasani, P. S. V. Gowtham, S. Yashwanth, S. S. Poorna, and A. K, "Speaker identification using MFCC feature extraction: A comparative study using GMM, CNN, RNN, KNN and random forest classifier," in *Proc. 2nd Int. Conf. Trends Electr., Electron., Comput. Eng. (TEECCON)*, 2023, pp. 287–292, DOI: 10.1109/TEECCON59234.2023.10335892 IEEE Xplore.

[6] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in Interspeech 2017, Aug. 2017, pp. 2616–2620. DOI: 10.21437/Interspeech.2017-950, arXiv.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964, IEEEX Xplore.

[8] . Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and V. Zue, "TIMIT Acoustic-phonetic Continuous Speech Corpus," Linguistic Data Consortium, Nov. 1992, LDT.

[9] A. Canavan and G. Zipperlen, "CALLFRIEND American English-Non-Southern Dialect," Linguistic Data Consortium, 1996. DOI: 10.35111/D37S-C536, LDC.

[10] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in Interspeech 2015, Sep. 2015, pp. 3586–3589. DOI: 10.21437/Interspeech.2015-711, ISCA.

[11] M. Hebert, "Text-Dependent Speaker Recognition," in Springer Handbook of Speech Processing, J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds., Berlin, Heidelberg: Springer, 2008, pp. 743–762. DOI: 10.1007/978-3-540-49127-9_37, Springer Link.

[12] S. A. El-Moneim et al., "Text-dependent and text-independent speaker recognition of reverberant speech based on CNN," International Journal of Speech Technology, vol. 24, no. 4, pp. 993–1006, Dec. 2021. DOI: 10.1007/s10772-021-09805-3, Springer Link.

[13] M. K. Singh, "Speaker Identification Using MFCC Feature Extraction ANN Classification Technique," Wireless Personal Communications, vol. 136, no. 1, pp. 453–467, May 2024. DOI: 10.1007/s11277-024-11282-1, Springer Link.

[14] J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A tutorial review," IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 74–99, Nov. 2015. DOI: 10.1109/MSP.2015.2462851, IEEEX Xplore.

[15] B. Dhonde and S. M. Jagade, "Feature Extraction Techniques in Speaker Recognition: A Review", International Journal on Recent Technologies in Mechanical and Electrical Engineering, vol. 2, no. 5, pp. 104-106, IJRMEE.

[16] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech Communication, vol. 52, no. 1, pp. 12–40, Jan. 2010. DOI: 10.1016/j.specom.2009.08.009, ELSEVIER.

[17] Y. Astuti, R. Hidayat, and A. Bejo, "Comparison of Feature Extraction for Speaker Identification System," in 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Dec. 2020, pp. 642–645. DOI: 10.1109/ISRITI51436.2020.9315332, IEEEX Xplore.

[18] P. Foggia, A. Greco, A. Roberto, A. Saggese, and M. Vento, "Identity, Gender, Age, and Emotion Recognition from Speaker Voice with Multi-task Deep Networks for Cognitive Robotics," Cognitive Computation, vol.

16, no. 5, pp. 2713–2723, Sep. 2024. DOI: 10.1007/s12559-023-10241-5, Springer Link.

[19] S. Song, S. Zhang, B. Schuller, L. Shen, and M. Valstar, “Noise Invariant Frame Selection: A Simple Method to Address the Background Noise Problem for Text-independent Speaker Verification” arXiv preprint, no. 1805.01259, May 2018. DOI: 10.48550/arXiv.1805.01259, arXiv.