Interspeech 2018
2-6 September 2018, Hyderabad

Spoofing Detection Using Adaptive Weighting Framework and Clustering Analysis

Yuanjun Zhao, Roberto Togneri, Victor Sreeram

School of Electrical, Electronic and Computer Engineering, The University of Western Australia (UWA), Australia

yuanjun.zhao@research.uwa.edu.au, (roberto.togneri, victor.sreeram)@uwa.edu.au

Abstract

Security of Automatic Speaker Verification (ASV) systems against imposters are now focusing on anti-spoofing countermeasures. Under the severe threat of various speech spoofing techniques, ASV systems can easily be 'fooled' by spoofed speech which sounds as real as human-beings. As two effective solutions, the Constant Q Cepstral Coefficients (CQCC) and the Scattering Cepstral Coefficients (SCC) perform well on the detection of artificial speech signals, especially for attacks from speech synthesis (SS) and voice conversion (VC). However, for spoofing subsets generated by different approaches, a low Equal Error Rate (EER) cannot be maintained. In this paper, an adaptive weighting based standalone detector is proposed to address the selective detection degradation. The clustering property of the genuine and the spoofed subsets are analysed for the selection of suitable weighting factors. With a Gaussian Mixture Model (GMM) classifier as the back-end, the proposed detector is evaluated on the ASVspoof 2015 database. The EERs of 0.01% and 0.20% are obtained on the *known* and the *unknown* attacks, respectively. This presents an essential complementation between the CQCC and the SCC and also promotes the future research on generalized countermeasures.

Index Terms: automatic speaker verification, anti-spoofing countermeasures, CQCC, SCC, adaptive weighting, clustering

1. Introduction

The Automatic Speaker Verification (ASV) systems take the task of examining the authenticity of any claimed identity [1]. Several commercial applications of ASV have been applied to security check platforms such as the bank e-transaction and the physical access control [2]. Recently, besides the low-cost and flexibility of ASV systems, there has been a significant increase in the level of unauthorized spoofed speech attacks [3]. Generally, four kinds of spoofing techniques are employed by imposters: impersonation, replay, speech synthesis (SS) and voice conversion (VC) [4].

In [5], the vulnerability of ASV systems to spoofing attacks is assessed and two essential approaches for anti-spoofing are introduced. One choice is to adjust the fundamental framework of ASV systems with a more robust mechanism. While another way is to apply dedicated external anti-spoofing countermeasures, especially ones with more generalized detection capability. The negative impacts caused by various spoofing attacks can be effectively reduced and, probably, eliminated.

Currently, in response to the risk of spoofing attacks, many researchers have begun to develop useful methods for anti-spoofing. In 2015, the first online ASV spoofing and countermeasures challenge was held. This special session focused on offering valid solutions to defend against the synthetic speech, including both speech synthesis and voice conversion [3].

In our previous work, the compressed sensing (CS) [6] framework was combined with a high-dimensional (HD) feature [7] and the proposed CS-HD features gained an Equal Error Rate (EER) of 0.01% for the *known* attacks [8]. Inspired by the combination of the cochlear filter cepstral coefficients and change in instantaneous frequency (CFCCIF) [9], which is the winning submission of the first challenge, a feature named Constant Q Cepstral Coefficients (CQCC) was proposed and an EER of 0.26% on the evaluation set was achieved [10]. The promising achievement of the CFCCIF and the CQCC had shown the importance of the filter-banks and the resampling process that were applied. In [11], a hierarchical scattering decomposition was employed to produce another speech feature referred to as Scattering Cepstral Coefficients (SCC). The first level coefficients of the SCC performed equivalently to the CQCC and the classical Mel-Frequency Cepstral Coefficients (MFCC). With this feature, the performance metric EER of the *known* subsets in the evaluation set reached as low as 0.03%.

However, released results in [11] demonstrates that it is hard for the CQCC and the SCC to perform an undifferentiated detection across all the spoofing subsets. The performance of specific subsets degrades with varying levels and this leads to an unstable anti-spoofing efficiency. For example, the detection performance of CQCC on the subset S10 is relatively better than SCC. The subset S10 is generated from the unit-selection based speech synthesis. The dynamic coefficients of CQCC can capture the artifacts caused by unnatural boundaries between units in a speech signal. While the analogous counterpart to dynamic coefficients are not implemented for SCC, which leads to a subdued detection performance on the subset S10. Conversely, for the other subsets, the SCC possesses a stable and outstanding detection level. This is because that the numerous orders of SCC can cover the full-scale of the frequency spectrum. Artificial details in the low and high frequency bands can be discovered by the SCC. The most significant contribution of this paper is thus the proposal of an adaptive weighting framework with an entirely new clustering approach to exploit the complementary behavior of CQCC and SCC.

In the proposed framework, the weighting factors are estimated by the evaluations on all the development subsets. By a voting process, the pairs of weighting factors and clusters are united. After that, the *known* and *unknown* speech signals in the evaluation set can be mapped into the labeled clusters. Then the allocated weighting factors are used afterwards for the scores calculation in the following GMM back-end.

The rest of the paper is organized as follow. A brief introduction to the CQCC and the SCC features is provided in section 2. The detailed concepts and framework are described in section 3. In section 4, the experimental results and relevant discussion are given. Finally, a conclusion of this paper and possible future works are detailed in section 5.

10.21437/Interspeech.2018-1042

2. Adaptive weighting framework and clustering approach

Herein the motivation for the application of the weighting framework is described. The discussion starts with a treatment of original CQCCs and SCCs before the introduction to the proposed adaptive weighting scheme and the clustering approach.

2.1. CQCC and SCC complementary features

In [10], the authors analyzed the effectiveness of the time-frequency representation and the constant Q transform to speech and music signals. Comparing with the conventional cepstral analysis, a conversion from geometric series space to linear space was adopted to match the equal-tempered scale corresponded to a bin spacing. The orthogonality of the discrete cosine transform (DCT) basis was preserved by performing a linearisation of the frequency scale of the constant Q transform. The extraction of the CQCC applied a polyphase anti-aliasing filter and a spline interpolation to complete the resampling of each signal with a uniform sampling rate. Actually, CQCCs are extracted in a traditional manner like MFCCs but with a new uniform resampling process. Due to the adjustment of the frequency resampling, the CQCC achieves an impressive performance on the ASVspoof 2015 database.

While in [11], the scattering spectrum [12] with the same nature as cochlear filters was used for the creation of the SCC. In addition to that, the close relation to the modulation spectra also contributed to enhance the spoofing detection ability. With a hierarchical structure based on wavelet filter-banks and modulus operators, raw speech signals were transferred to scalograms of different depth. Then scattering coefficients at each level can be estimated by windowing and computing the average value. Logarithms of the scattering coefficients from several levels were concatenated to build a feature vector. By taking a DCT of this vector, the SCC was eventually collected.

From the detection results of the evaluation set in [10] [11], it is shown that a lower EER can be reached on the subset S10 compared to the SCC which, conversely, takes the lead to the other nine subsets. Specifically, the CQCC captures more information of artifacts hidden in unit boundaries than the SCC and this brings on a significantly decreased EER in detecting unit-selection based spoofing, S10. For S10 using CQCC the EER is 1.07% to be compared to SCC with an EER of 3.94% [11]. By contrast, an analysis of the F-ratios for the first three levels of the SCC reveals that the forged components in both the low frequency and the high frequency region are discriminative. When the first several levels are used in conjunction, the accuracy of the detection increases greatly. For the subsets S1 to S9, the EER is 0.02% with SCC compared to CQCC with EER of 0.17% [11].

These results confirm the complementary performance of CQCC and SCC and present the question as to whether these can be successfully fused. We propose an approach to do just this as described in the next section.

2.2. Proposed framework

In this work, the dataset used is the ASVspoof 2015 Corpus [3]. The detailed information of this database is shown in Table 1. Three sets are included consisting of genuine (human) speech and spoofed speech. No channel or background noise effects are added. There are 10 types of spoofing attack, namely S1 to S10. In all three speech sets, the S1-S5 are set as *known* attacks and the S6-S10 are the *unknown* attacks. Note that the spoofed speech in the training and development sets only consist of the five *known* attacks. While the spoofed speech in the evaluation set are made up of unseen *known* attack speech data and five *unknown* attack speech data. The database protocol motivates the design of a clustering method for the task of weighting factors assignment. The training and the development sets are prepared for training and tuning the detector while the evaluation set is for testing.

Table 1: *Statistics of the ASVspoof 2015 database*

Subset	#Speaker		#Utterances	
	Male	Female	Genuine	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	$\approx$ 200000

Figure 1 illustrates the scheme of the proposed adaptive weighting framework. The CQCCs and the SCCs of the raw speech signals from all the sets in ASVspoof 2015 corpus are extracted with the same settings in [10] [11]. Note that the voice activity detection (VAD) is discarded here to take the silent regions into consideration. In the step of weighting factors selection, the training and development sets are utilized for the tuning of weighting factors α_i $(i = 1, 2, \ldots, 5)$ for five spoofing categories and factor α_0 for the human subset provided in the development protocol, thus we have six weighting factors. The weighting formula is shown as below:

$$\text{final\_score} = \alpha_i \cdot \text{score}_{\text{CQCC}} + (1 - \alpha_i) \cdot \text{score}_{\text{SCC}} \quad (1)$$

where the scores are represented by the log-likelihood ratio (LLR). For all the subsets in the development set, we search over all the available values of weighting factors α_i ranging from 0 to 1. When the EER is minimized by a certain weighted score, the associated weighting factor is fixed for that subset.

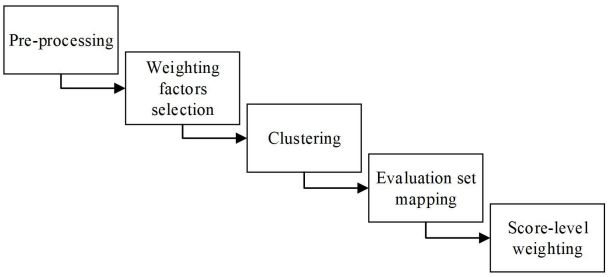

Figure 1: *Scheme of the proposed adaptive weighting framework.*

To arrange a weighting factor for every speech in the evaluation set without any prior knowledge, a new clustering method is proposed in this paper. The flowchart of the clustering process is shown in Figure 2. After the extraction of the CQCC, the cepstral mean and variance normalization (CMVN) is applied. Then two universal background models (UBM) are trained for human and spoofed speech respectively. Following this, the GMM mean supervectors are estimated by the EM algorithm. To reduce the computation complexity and acquire a more compact representation, the locality preserving projections (LPP) [13] is employed. The newly generated short vectors are named

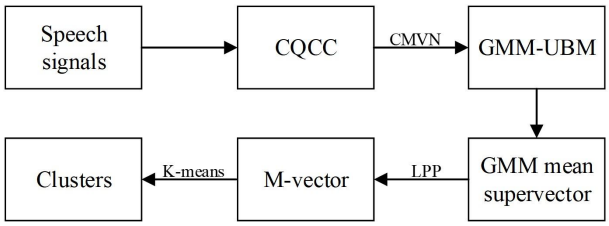

Figure 2: *Clustering process flowchart.*

the M-vectors and are followed by the K-means algorithm for building clusters.

With a reasonable set of parameters, a total of six clusters c_i $(i = 0, 1, \ldots, 5)$ are created with the speech signals from the training and the development sets. After a voting operation, each cluster is labeled as one of the five known attacks or genuine speech based on the dominant speech representing that cluster. The labeled clusters are denoted as c_{human} and $c_{s1}, c_{s2}, \ldots, c_{s5}$. All the six weighting factors and clusters are matched as pairs $(\alpha_0, c_{human}), (\alpha_1, c_{s1}), \ldots, (\alpha_5, c_{s5})$. The centroids of six clusters are recorded for the subsequent assignment of the new data point from the evaluation set. By calculating the distance to each cluster center, every new speech from the evaluation subsets is enrolled to an existing cluster. Consequently, the mapping from 11 subsets (one for the genuine subset and 10 for subsets S1-S10) to 6 trained clusters (c_{human} and $c_{s1}, c_{s2}, \ldots, c_{s5}$) are realized. Therefore, each undetected signal in the evaluation set is allocated with a weighting factor used for the scores calculation. All the final-scores are assembled to estimate the detection EERs.

In this framework, the weighting factors are assigned adaptively with no need for the prior knowledge of the evaluation set. This is based on the clustering process and the hypothesis that the latent artifacts across different sets are with similar natures and traits. The experimental results and relevant analysis are given in next section to assess the validity of the proposed framework.

3. Experimental results

3.1. Experiments settings

3.1.1. Feature extraction

The original CQCC features are extracted by an open-source MATLAB toolkit (http://audio.eurecom.fr/content/software). The parameters are the same as the configuration in [10]. The maximum and the minimum frequency in the constant Q transform are set as $F_{\max} = F_{sample}/2$ and $F_{\min} = F_{\max}/2^9$ respectively. The Nyquist frequency of the database is $F_{sample} = 16$kHz. The number of octaves is 9 and the number of bins per octave B is set to 96, which results in a time shift of 8 ms. Parameter γ is set to $\gamma = \Gamma = 228.7 * \left(2^{(1/B)} - 2^{(-1/B)}\right)$. $d = 16$ is the re-sampling period. The dimension of the CQCC static coefficients is set to 19 with appended C_0 which makes its length is 20. Acceleration coefficients, namely delta-delta ($\Delta\Delta$), are calculated and used in isolation. Experiments later are performed with only the CQCC acceleration coefficients (CQCC-A).

The extraction of the scattering coefficients is completed with the publicly available toolbox Scattering (http://www.di.ens.fr/data/scattering/). The first three levels of coefficients are concatenated together. A DCT is performed and the first 60 coefficients are retained as the final scattering cepstral coefficients. The voice activity detector applied in [11] is discarded in this work.

3.1.2. Parameter settings

In the clustering stage, the GMM-UBM system is from the MSR Identity Toolbox v1.0 [14]. All the GMMs in this work utilize 512 mixture components. The LPP is implemented by the open-source Matlab Toolbox for Dimensionality Reduction (https://lvdmaaten.github.io/drtoolbox/). The dimension of the created vectors followed by k-means clustering algorithm is set to 20.

All the scores of the detector are represented by the Log-likelihood ratio (LLR). The EER is defined as the operating point on the Detection error tradeoff (DET) curve, where the false acceptance rate (FAR) is equal to the false rejection rate (FRR). A lower EER indicates a better detection performance.

3.2. Results and analysis

Figure 3 gives the clusters distribution of the evaluation set in the ASVspoof 2015 corpus and Figure 4 demonstrates the relationships of the six labeled clusters. These visualizations are produced by the t-distributed Stochastic Neighbor (t-SNE) embedding algorithm applied onto the extracted M-vector of each utterance.

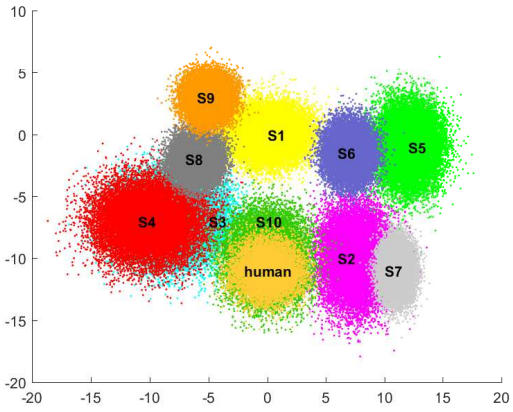

Figure 3: *Cluster distribution of the evaluation set*

Comparing Figure 3 with Figure 4, it is obvious that there is a correlation of the five *unknown* attacks with the *known* attacks. It is also noted that S3 and S4 are overlapped in most parts because they are generated by the same spoofing technique with different amounts of data. And the subset S10 appears to overlap the vast majority of the human subset. Spoofed speech of subset S10 are fabricated by the unit-selection synthesis. The sub-words units are from databases of natural speech. This is the main reason that most published spoofing detector systems perform poorly in differentiating S10 from the human subset. The result of the mapping in Figure 4 confirms our assumption in section 2. That is, the spoofed speech in the evaluation set across both *known* and *unknown* attacks contain similar information to the *known* attacks in the training and development sets. Given this condition, they can be re-clustered into the existing labeled clusters (e.g. *known* S2 and *unknown* S7 map to *known* S2).

Table 2: *The EERs(%) of the development set and the evaluation set of the ASVspoof 2015 corpus.*

Feature	Development set						Evaluation set													
	S1	S2	S3	S4	S5	Ave.	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Known	Unknown	S1-S9	Ave.
CQCC-A [10]	0.01	0.17	0.00	0.00	0.12	0.06	0.01	0.11	0.00	0.00	0.13	0.10	0.06	1.03	0.05	1.07	0.05	0.46	0.17	0.26
SCC [11]	0.00	0.09	0.00	0.00	0.27	0.07	0.01	0.12	0.00	0.00	0.02	0.01	0.01	0.03	0.01	3.94	0.03	0.80	0.02	0.42
Proposed	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00	0.01	0.00	0.00	0.02	0.01	0.01	0.00	0.00	0.95	**0.01**	**0.19**	**0.01**	**0.10**

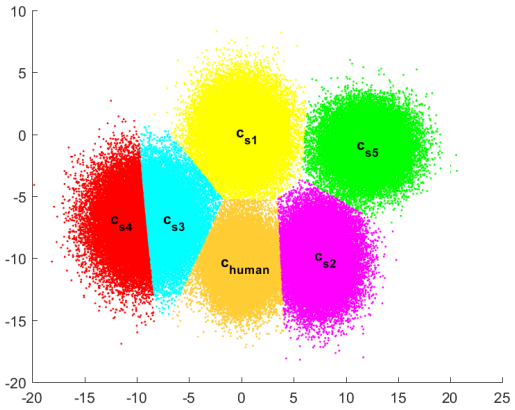

Figure 4: *Visualization of the six labeled clusters*

The testing results on the development set are given in Table 2. The CQCC-A is defined as the CQCC acceleration coefficients. From the table it can be seen that the proposed adaptive weighting scheme efficiently improves the detection accuracy. And this also implies the adaptive selection of the weighting factors is guaranteed.

Table 2 also shows the detection performance of the CQCC, the SCC and our proposed score-level weighting framework. The EERs of all the ten spoofing subsets are listed, appended with averaged results of the *known*, the *unknown*, the subsets S1 to S9 and the whole evaluation set. By comparison, all the EERs of the proposed method are comparable or better than those of the original features. The EERs of S1 to S9 are at a stable low level ranging from 0.00% to 0.02%. It is worth noting that the EER of the subset S10 is 0.95% which is an improvement of 75.9% over the SCC feature and 11.2% over the CQCC feature. This results in an impressive advancement on the average performance with an overall EER of 0.10% which is 61.5% better than the competing CQCC.

The selected weighting factors are listed in Table 3. For clusters c_{s3} and c_{s4}, the weighting factors can be any value ranging from 0.00 to 1.00. This is because the original CQCC and SCC features have already had a qualified detection performance on these two subsets. Meanwhile, for cluster c_{human} the range of optional weighting factors is limited. It typically suggests a more elaborate weight allocation because of the different detection capability of the original CQCC and SCC.

A generalized anti-spoofing countermeasure will be in great demand for practical scenarios. The effectiveness of the proposed adaptive weighting scheme can be treated as one of the possible solutions to the generalization issue. Furthermore, due to uncertain varieties and mixtures of spoofing attacks in real-world environment, a generalized detector with only one anti-spoofing strategy does not usually work consistently. The countermeasures based on weighting or ensemble methods are encouraging ways to develop such generalized spoofing detectors. Moreover, any extra data produced by new spoofing methods based on Deep Learning (DL) should be supplemented into the corpus to improve the cluster generation and weighting factors.

Table 3: *List of the selected weighting factors*

Clusters	Weighting factors
c_{human}	$\alpha_0 = 0.54 \sim 0.56$
c_{s1}	$\alpha_1 = 0.00 \sim 0.96$
c_{s2}	$\alpha_2 = 0.43 \sim 0.53$
c_{s3}	$\alpha_3 = 0.00 \sim 1.00$
c_{s4}	$\alpha_4 = 0.00 \sim 1.00$
c_{s5}	$\alpha_5 = 0.66 \sim 0.73$

4. Conclusions

In this paper, a novel adaptive weighting framework for score-level fusion of the spoofing detector is proposed. A new clustering method is also introduced for the analysis of the data structure. To address the degradation problem on individual speech subsets, the original CQCC and SCC features are merged at the score level. The weighting factors are chosen on the basis of the clustering distribution of genuine and spoofed signals. Our proposed weighting framework has been shown to provide a lower overall EER on the ASVspoof 2015 database compared to either CQCC or SCC alone. Furthermore, the analysis of the experiments in this work also illustrates that a systematic integration of diverse anti-spoofing methods is a promising strategy for a more generalized countermeasure. Future work may be focused on exploiting more potential combinations of anti-spoofing applications for speaker recognition and relevant areas.

5. Acknowledgements

This work was supported by the International Postgraduate Research Scholarship (IPRS) from the University of Western Australia.

6. References

[1] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23–61, Secondquarter 2011.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, p. 101962, Apr 2004.

[3] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, June 2017.

[4] Z. Wu, P. L. D. Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z. H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768–783, April 2016.

[5] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130 – 153, 2015.

[6] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[7] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2119–2123.

[8] Y. Zhao, R. Togneri, and V. Sreeram, "Compressed high dimensional features for speaker spoofing detection," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec 2017, pp. 569–572.

[9] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech and Language*, vol. 45, pp. 516–535, Sep 2017.

[11] K. Sriskandaraja, V. Sethu, E. Ambikairajah, and H. Li, "Front-end for antispoofing countermeasures in speaker verification: Scattering spectral decomposition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 632–643, June 2017.

[12] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, Aug 2014.

[13] Y. Tang and R. Rose, "A study of using locality preserving projections for feature extraction in speech recognition," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 1569–1572.

[14] S. O. Sadjadi, M. Slaney, and L. Heck, "Msr identity toolbox v1. 0: A matlab toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, vol. 1, no. 4, pp. 1–32, 2013.