INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Experimental analysis of features for replay attack detection–Results on the ASVspoof 2017 Challenge

Roberto Font[1], *Juan M. Espín*[1], *María José Cano*[1]

[1]Biometric Vox S. L., Spain

{roberto.font,jm.espin,mariajose.cano}@biometricvox.com

Abstract

This paper presents an experimental comparison of different features for the detection of replay spoofing attacks in Automatic Speaker Verification systems. We evaluate the proposed countermeasures using two recently introduced databases, including the dataset provided for the ASVspoof 2017 challenge. This challenge provides researchers with a common framework for the evaluation of replay attack detection systems, with a particular focus on the generalization to new, unknown conditions (for instance, replay devices different from those used during system training). Our cross–database experiments show that, although achieving this level of generalization is indeed a challenging task, it is possible to train classifiers that exhibit stable and consistent results across different experiments. The proposed approach for the ASVspoof 2017 challenge consists in the score–level fusion of several base classifiers using logistic regression. These base classifiers are 2–class Gaussian Mixture Models (GMMs) representing *genuine* and *spoofed* speech respectively. Our best system achieves an Equal Error Rate of 10.52% on the challenge evaluation set. As a result of this set of experiments, we provide some general conclusions regarding feature extraction for replay attack detection and identify which features show the most promising results.

Index Terms: Automatic speaker verification, anti-spoofing, ASVspoof 2017

1. Introduction

In the last decades, significant progress has been made in the field of speaker recognition. Indeed, this technology has reached a degree of maturity that allows widespread adoption in real–world applications. This, however, requires systems that are not only reliable and robust but also resilient to attacks. In this sense, several studies have highlighted the vulnerability of Automatic Speaker Verification (ASV) systems to spoofing attacks when no countermeasures are taken [1, 2, 3]. The main types of attacks that an ASV system can face are [2]: (1) impersonation, where the attacker mimics the voice of the legitimate speaker; (2) replay, in which the attacker presents a prerecorded voice sample of the legitimate speaker; (3) voice synthesis, where a text-to-speech system adapted to the characteristics of the legitimate speaker is used; and (4) voice conversion, where the speech signal of the attacker is altered to resemble that of the legitimate speaker.

Replay is the most accessible spoofing attack, needing no special signal processing knowledge. However, it has received little attention to date. This is partly due to the lack of publicly available databases and standardized benchmarks. Previous work [4, 5, 6] used small sized in-house databases collected with a small set of recording and playback devices, which makes it difficult to compare their results with those of other studies. Recently, some efforts have been devoted to overcome this difficulty, with the collection of databases like AVspoof [7] or RedDots Replayed [8]. In this vein, ASVspoof 2017 is the first initiative to provide a common framework with standard corpus, protocol and metrics.

The challenge task at ASVspoof 2017 is replay attack detection: given a test utterance, determine whether it is a genuine human voice (live recording) or a replay recording. The objective is to develop a detection system that performs consistently on challenging unseen conditions, for example, recording environments or playback devices different from those used to train the system.

In this paper we present an experimental study of different features for replay attack detection in Automatic Speaker Verification systems. A total of nine different features are considered, including cross–database experiments to assess how different systems perform on mismatched conditions (different devices, phonetic content, collection protocols, etc.). We have adopted a similar approach to that presented in [9, 10]. Up to the best knowledge of the authors, this is the first comprehensive study of this kind for replay spoofing attack detection.

We provide some general insights into feature extraction for replay attack detection and show how we developed our final submission for the ASVspoof 2017 challenge. Our approach uses as base classifiers two Gaussian Mixture Models (for *genuine* and *spoof* speech respectively) with the log–likelihood ratio as score. The final system is the score–level fusion of a subset of base systems using logistic regression. Our best individual system achieves an Equal Error Rate (EER) of 11.49% and our best overall result is EER$= 10.52\%$.

The rest of the paper is organized as follows. In Section 2 we briefly describe the databases used in our study. Experimental results are presented in Section 3. Finally, Section 4 presents our conclusions.

2. Databases

In this section we briefly describe the databases used in our experiments: the ASVspoof 2017 challenge dataset and the AVspoof corpus.

2.1. ASVspoof 2017

The ASVspoof 2017 challenge data is primarily based on the RedDots [11] database, as source of genuine recordings, and the RedDots Replayed [8] database as source of spoof replay recordings. The RedDots Replayed corpus was created by replaying Part 01 of the original corpus, which consists of 10 common short phrases, through a variety of recording environments and recording/playback devices.

Data for the challenge was distributed partitioned into three subsets: *training*, *development* and *evaluation*. Each speech file in *training* and *development* was labeled as genuine or spoof.

http://dx.doi.org/10.21437/Interspeech.2017-450

Information regarding phrase ID, speaker, recording environment, recording device and playback device was also available. For the *evaluation* subset, only the phrase ID was available.

The challenge comprises two different conditions: **common condition**, where only ASVspoof 2017 data can be used to train detection systems, and **flexible condition** where any external data can be used.

2.2. AVspoof

The AVspoof corpus [7] contains recordings of 44 speakers captured in 4 sessions using two different smart-phones and a laptop. AVspoof comprises a collection of spoofing attacks, divided into *logical access attacks* and *presentation attacks*. Logical access attacks include voice synthesis and voice conversion. Presentation attacks include: (1) direct replay, where genuine speech is played back using two different smart-phones and a laptop using both its built-in speakers and external high quality loudspeakers, (2) synthesized speech reproduced with a laptop, and (3) converted voice replayed using a laptop.

In our experiments we use the training and development splits created for the BTAS2016 Speaker Anti-spoofing Competition [12] but excluding all voice synthesis and voice conversion attacks.

3. Experiments

In this paper, we have focused our efforts on finding discriminative features. This approach is in line with the general observation (e.g., see [13]) that the design of spoofing countermeasures should start with the search for a good set of discriminative features rather than the design of complex classifiers. Therefore, our proposal uses relatively simple classifiers. In particular, two Gaussian Mixture Models (GMMs) are trained on genuine and spoofed speech respectively using Maximum Likelihood estimation. The score is computed as the log–likelihood ratio for the test utterance given both classifiers.

We have performed our comparative study on a set of features similar to those considered in [9, 10]: classifiers were trained using Constant-Q Cepstral Coefficients (CQCCs) [13], Mel Frequency Cepstral Coefficients (MFCCs) [14], Linear Frequency Cepstral Coefficients (LFCCs), Inverted Mel Frequency Cepstral Coefficients (IMFCCs) [15], Rectangular Filter Cepstral Coefficients (RFCCs) [16], Linear Prediction Cepstral Coefficients (LPCCs) [17], Subband Spectral Flux Coefficients (SSFCs) [18], Subband Spectral Centroid Frequency Coefficients (SCFCs) and Subband Spectral Centroid Magnitude Coefficients (SCMCs) [19, 20, 21].

Two sets of experiments are presented. In the first one we work with ASVspoof 2017 data only and our goal is to achieve a development set error as low as possible. In the second set of experiments, we perform cross-database experiments and focus on finding a configuration that shows consistent performance and overfits on a particular database as little as possible. In this case, our goal is not to obtain the best possible results on a particular test set, but to obtain consistent and stable performance across unseen conditions.

3.1. Experiments on ASVspoof 2017 database

In this first set of experiments, we focus on ASVspoof 2017 database, which corresponds with common condition in the ASVspoof 2017 challenge.

For model selection and hyperparameter tuning, we trained the classifiers on the training portion of the data and evaluated them on the development set. Table 1 shows the optimal set of parameters found for each system: the number of static coefficients (we append delta and double delta coefficients in all cases), the frequency range considered and the applied feature normalization technique. CQCCs are the only features that use Cepstral Mean Normalization (CMN). For the rest of the systems, adding feature normalization was found detrimental. None of the systems use any Voice Activity Detection, since removing non–speech frames showed to hurt performance.

It is worth noting that the number of static coefficients that was found optimal is much higher than in other applications (in speaker recognition, for example, 20 static coefficients are typically used). We discuss this issue in Section 3.3.

All GMM classifiers have 512 Gaussian components.

Table 1: *Parameters used for feature extraction*

Feature	Dimension	f_{min}–f_{max}	Normalization
CQCCs	$50 + \Delta + \Delta\Delta$	$15.62 - 8000$	CMN
MFCCs	$70 + \Delta + \Delta\Delta$	$300-8000$	–
LFCCs	$70 + \Delta + \Delta\Delta$	$100-7800$	–
IMFCCs	$60 + \Delta + \Delta\Delta$	$200-8000$	–
RFCCs	$30 + \Delta + \Delta\Delta$	$200-8000$	–
LPCCs	$50 + \Delta + \Delta\Delta$	–	–
SCFCs	$20 + \Delta + \Delta\Delta$	$100-8000$	–
SCMCs	$40 + \Delta + \Delta\Delta$	$100-8000$	–
SSFCs	$20 + \Delta + \Delta\Delta$	$100-8000$	–

The results obtained from the development set are shown in Table 2, second column. The metric considered is Equal Error Rate (EER), which is the official metric for the ASVspoof 2017 challenge. We can see that the best result, EER$= 3.85\%$, is achieved by the Inverted Mel Frequency Cepstral Coefficients system.

Table 2: *Results on ASVspoof 2017 database. Common condition. Best result for each set is shown in bold type.*

Feature	Development EER (%)	Evaluation EER (%)
Baseline	10.35	24.65
CQCCs	8.20	17.41
MFCCs	7.76	27.12
LFCCs	5.61	26.27
IMFCCs	**3.85**	30.91
RFCCs	6.91	11.90
LPCCs	5.94	25.20
SCFCs	24.51	24.83
SCMCs	9.32	**11.49**
SSFCs	12.81	22.38
Fusion I	–	17.62
Fusion II	–	14.37

3.2. Submission to ASVspoof 2017 challenge. Common condition

Once we had the optimal set of parameters for each feature, and a set of development set scores for each system, we trained new classifiers by pooling training and development sets and used these new systems to score the evaluation set. To generate our final submission, we fused the scores of different classifiers using logistic regression. The fusion procedure is as follows:

- Development set scores were generated using classifiers trained on the training subset.
- Evaluation set scores were generated using the final classifiers trained on pooled training/development data.
- Phrase ID information was one–hot encoded and added as feature.
- A logistic regression classifier was trained on development scores + one–hot encoded phrase information.
- This classifier was used to produce final evaluation scores.

To tune the logistic regression classifier and choose the best possible subset of scores, we used *leave–one–label–out* cross–validation, where labels can be *recording device*, *playback device*, *environment*, or each unique combination of the above. We used playback device since it seemed to produce the most consistent results.

We trained two systems:

Fusion I which is our *primary* submission uses the best subset of scores found by the described cross–validation procedure: CQCCs + IMFCCs + SCMCs + Phrase ID.

Fusion II uses the best subset of scores excluding the best individual system (IMFCCs): CQCCs + LPCCs + RFCCs + Phrase ID.

Results are summarized in Table 2. The baseline system provided by the organization based on CQCCs is included as a reference. As we can see, evaluation was a very challenging set and the performance of most systems degrades considerably compared to the development scores. However, the score fusion strategy helped to alleviate the impact of the overfitting suffered by some individual systems. It is worth noting that Subband Spectral Centroid Magnitude Coefficients showed a remarkably stable behavior, with the lower degradation and the best overall evaluation error (EER$= 11.49\%$).

3.3. Cross-database experiments

In this set of experiments, we train the classifiers using ASVspoof 2017 training set and evaluate on BTAS 2016 AVspoof development set and vice versa. The aim is to find classifiers that generalize well to new unseen conditions. The set of parameters for feature extraction that we found optimal in this cross-database setup is summarized in Table 3. In this case, all classifiers have 128 Gaussian components.

We can see that, in all cases, CMN was beneficial. However, using more complex feature normalization techniques, like Cepstral Mean and Variance Normalization or Feature Warping [22], degraded performance.

The optimal number of static coefficients is lower than in previous experiments, which suggests that previous choice was indeed overfitting on the development set, but still remarkably high, in the range of 20–40 (up to 60 in the case of LFCCs). This might suggest that capturing the subtle effect of a playback device requires a fine spectral detail.

Results are summarized in Table 4. We can see that cross–database error rates are much higher than those obtained within the same database. This highlights the importance of developing robust countermeasures that generalize well to new unseen attacks. Results show that the performance of some systems, mainly RFCC, varies greatly from one database to another, while in other systems, like CQCC or SCMC, is quite stable (Table 4, third and fourth columns).

Table 3: *Parameters used for feature extraction*

Feature	Dimension	f_{min}–f_{max}	Normalization
CQCCs	$30 + \Delta + \Delta\Delta$	$15.62 - 8000$	CMN
MFCCs	$30 + \Delta + \Delta\Delta$	$300-8000$	CMN
LFCCs	$60 + \Delta + \Delta\Delta$	$0-8000$	CMN
IMFCCs	$40 + \Delta + \Delta\Delta$	$200-8000$	CMN
RFCCs	$20 + \Delta + \Delta\Delta$	$200-8000$	CMN
LPCCs	$30 + \Delta + \Delta\Delta$	$-$	CMN
SCFCs	$20 + \Delta + \Delta\Delta$	$100-8000$	CMN
SCMCs	$40 + \Delta + \Delta\Delta$	$100-8000$	CMN
SSFCs	$20 + \Delta + \Delta\Delta$	$100-8000$	CMN

3.4. Submission to ASVspoof 2017 challenge. Flexible condition

Taking into account the above results and parameter selection, we trained new systems on pooled ASVspoof 2017 train and development data, and used these systems to score the evaluation trials. Note that, although the *flexible* condition allowed the use of external data to train the systems, we trained on ASVspoof 2017 data only, and used external data for the parameter selection. Figure 1 shows the DET curve for a subset of systems evaluated on ASVspoof 2017 development set.

Regarding the system fusion, we followed the same procedure described in Section 3.2. In this case, the best system combination was RFCCs + LFCCs with no phrase information. This was our *primary* submission with an EER=10.52%, our best overall result.

Table 5: *Results on ASVspoof 2017 database. Flexible condition. (Second column is copied from Table 4 for easier comparison.*

Feature	Development EER (%)	Evaluation EER (%)
CQCCs	9.85	17.43
MFCCs	20.89	26.13
LFCCs	10.31	16.54
IMFCCs	12.75	18.85
RFCCs	**8.35**	17.73
LPCCs	10.70	16.45
SCFCs	25.48	25.03
SCMCs	11.62	14.35
SSFCs	12.55	21.88
RFCC + LFCC	–	**10.52**

Comparing results in Table 5 with those in Table 4, we can see that ASVspoof 2017 evaluation error rates are very consistent with those found in our cross-database experiments. Another result that strikes our attention is the consistent performance of the SCMC system, which is again the best individual system and the one that presents the lower degradation from development to evaluation score.

4. Conclusions

We have presented results from an experimental comparison of different features for replay spoofing attack detection in Automatic Speaker Verification systems. We have conducted our experiments using two recently published databases focusing on finding configurations that provide stable performance against

Table 4: *Experiment results using ASVspoof 2017 and AVspoof databases. First row is the dataset used to train models, second row is the test set. Third and fourth columns correspond to cross–database results.*

Feature	BTAS 2016 Train		ASVspoof 2017 Train	
	BTAS 2016 AVspoof Development EER (%)	ASVspoof 2017 Development EER (%)	BTAS 2016 AVspoof Development EER (%)	ASVspoof 2017 Development EER (%)
CQCCs	4.34	18.17	22.18	9.85
MFCCs	6.01	29.95	27.95	20.89
LFCCs	1.54	27.45	17.29	10.31
IMFCCs	4.38	27.25	28.33	12.75
RFCCs	12.49	18.50	28.57	8.35
LPCCs	1.09	22.81	13.88	10.70
SCFCs	15.73	49.35	48.30	25.48
SCMCs	10.29	19.24	17.69	11.62
SSFCs	9.07	20.72	28.51	12.55

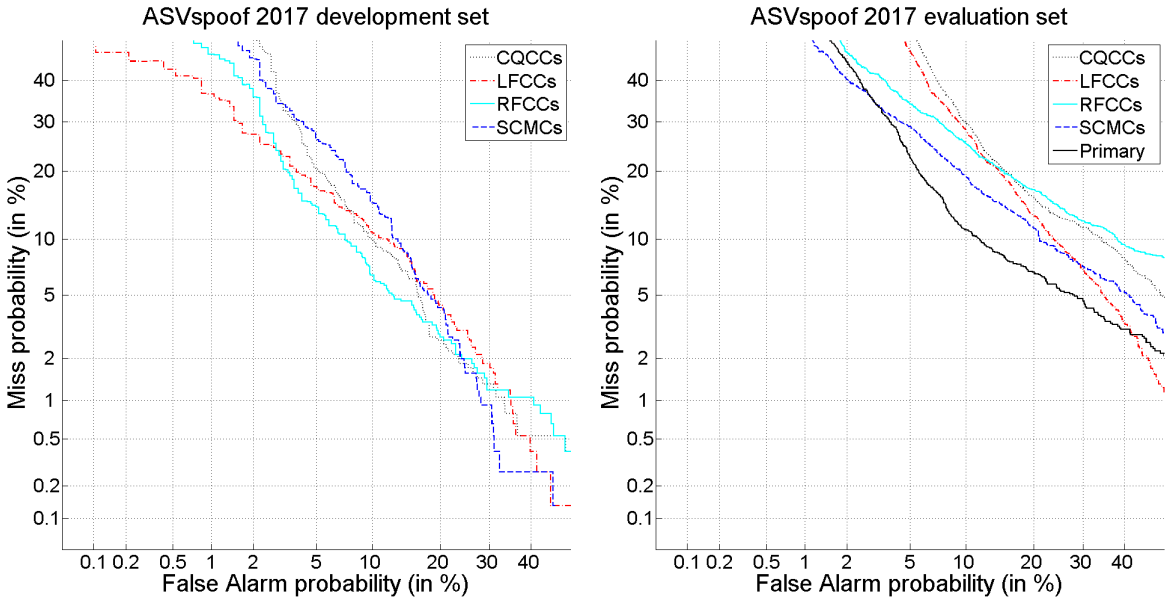

Figure 1: *DET curves for ASVspoof 2017 development (left) and evaluation (right) sets.*

new unknown attacks, different from those used to train the classifier.

This experimental study has allowed us to extract some general guidelines regarding feature extraction for replay attack detection:

- The use of voice activity detection to remove silence frames seemed to hurt performance in all cases. This might suggest that there is some information about playback device in non–speech frames.
- Cepstral Mean Normalization showed to improve generalization. However, more advanced techniques like CMVN, sliding–window CMVN, or Feature Warping were not beneficial.
- A number of filters and static coefficients higher than usual in other applications seems to improve detection accuracy.
- Using static plus delta plus delta–delta coefficients provided the best results in all cases.
- Subband Spectral Centroid Magnitude Coefficients (SCMCs) seem to be very promising features for replay attack detection. These systems showed both the lower error rates and the most consistent performance across different experiments.

We used the above insights to develop our submission to the ASVspoof 2017 Automatic Speaker Verification Spoofing and Countermeasures Challenge. The final system is the score level fusion of several subsystems using logistic regression. Our best system achieved an EER=10.52%.

5. Acknowledgements

The authors would like to thank the organizers of ASVspoof 2017 challenge.

6. References

[1] P. D. Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280–2290, 2012.

[2] 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, no. 0, pp. 130–153, 2015.

[3] Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *APSIPA ASC*, 2012.

[4] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *International Conference of the Biometrics Special Interest Group BIOSIG*, 2014.

[5] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Biometrics and ID Management*, C. Vielhauer, J. Dittmann, A. Drygajlo, N. Juul, and M. Fairhurst, Eds. Springer, 2011, pp. 274–285.

[6] Z. Wang, G. Wei, and Q. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *IEEE Int. Conf. Machine Learning and Cybernetics (ICMLC)*, 2011.

[7] S. K. Ergnay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2015, pp. 1–6.

[8] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z. H. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamäki, and K. A. Lee, "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *42nd IEEE Internation Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[9] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany, Sep. 2015, pp. 2087–2091.

[10] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, CA, USA, Sep. 2016.

[11] K. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, M. J. Alam, A. Swart, and J. Perez, "The reddots data collection for speaker recognition," in *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[12] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonalves, A. G. S. Mello, R. P. V. Violato, F. O. Simes, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi, H. Dinkel, N. Chen, Y. Qian, G. S. D. Paul, and M. Sahidullah, "Overview of BTAS 2016 speaker anti-spoofing competition," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2016.

[13] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech and Language*, to appear.

[14] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[15] S. Chakroborty, A. Roy, and G. Saha, "Improved closed set text-independent speaker verification by combining MFCC with evidence from flipped filter banks," *International Journal of Signal Processing*, vol. 4, no. 2, pp. 114–122, 2007.

[16] T. hasan, S. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. Hansen, "CRSS systems for 2012 NIST speaker recognition evaluation," in *ICASSP 2013 – International Conference on Acoustics, Speech and Signal Processing*, 2013.

[17] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[18] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997, pp. 1331–1334.

[19] E. A. P. N. Le, J. Epps, V. Sethu, , and E. Choi, "Investigation of spectral centroid features for cognitive load classification," *Speech Communication*, vol. 53, no. 4, pp. 540–551, 2011.

[20] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition," in *A Speaker Odyssey - The Speaker Recognition Workshop*, 2010.

[21] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998, pp. 617–620.

[22] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey - The Speaker Recognition Workshop*, 2001.