INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Investigation of Sub-Band Discriminative Information between Spoofed and Genuine Speech

Kaavya Sriskandaraja[1,2], *Vidhyasaharan Sethu*[1], *Phu Ngoc Le*[1,2], *Eliathamby Ambikairajah*[1,2]

[1] School of Electrical Engineering and Telecommunications, UNSW, Australia
[2] ATP Research Laboratory, National ICT Australia (NICTA), Australia

k.sriskandaraja@unsw.edu.au, v.sethu@unsw.edu.au, ngoc.le@unsw.edu.au, e.ambikairajah@unsw.edu.au

Abstract

A speaker verification system should include effective precautions against malicious spoofing attacks, and although some initial countermeasures have been recently proposed, this remains a challenging research problem. This paper investigates discrimination between spoofed and genuine speech, as a function of frequency bands, across the speech bandwidth. Findings from our investigation inform some proposed filter bank design approaches for discrimination of spoofed speech. Experiments are conducted on the Spoofing and Anti-Spoofing (SAS) corpus using the proposed frequency-selective approach demonstrates an 11% relative improvement in terms of equal error rate compared with a conventional mel filter bank.

Index Terms: speech recognition, automatic speaker verification, spoofing and anti-spoofing, voice conversion, SAS, speech synthetic computational paralinguistic

1. Introduction

Automatic speaker verification (ASV) is a non-invasive and convenient biometric person authentication technique, which aims to recognise people by analysing their speech. This has become more mature and has been deployed in commercial person authentication applications, like credit card verification, e-commerce, telephone based secure access in call centres, secure building access and suspect identification. Generally, speaker verification is used in telephone based systems where there is no face to face contact. Therefore, speech is more susceptible to spoofing attacks than other biometric signals [1-3]. This can lead to security concerns when deploying ASV systems and the development of techniques for the detection and prevention of these spoofing attacks is a critical area of research.

Spoofing attacks can be categorized as impersonation [3], replay [4], voice conversion [2] and speech synthesis [5, 6]. Among these, impersonations are generally not very effective or practical for large scale attacks. Replay attacks are the most accessible and can be highly effective, but they can be successfully addressed using different types of precautions, such as prompting the user for different pass-phrases each time the authentication process is employed. Speech synthesis (SS) and voice conversion (VC) have been identified as the most accessible approaches for attackers because of the availability of VC and SS toolboxes and their ability to be employed even when different pass-phrases are prompted by the ASV system. We focus on voice conversion and speech synthesis based spoofing attacks in this paper.

There are two broad approaches to address VC and SS based spoofing attacks [6]. One is to further improve the accuracy of the speaker verification system in order to directly detect spoofing [7]. The other one is to design a specific countermeasure, which should make an independent decision in parallel to the speaker verification system [8, 9]. Although both methods have their merits, the standalone spoofing detection approach can be easily integrated into existing ASV systems.

While, a number of countermeasures to spoofing have been proposed recently, most of these make use of prior knowledge about the specific VC and/or SS method employed in the spoofing attack. For example, discriminative features like higher order Mel-cepstral coefficients [10], which are based on the knowledge of HMM-based speech synthesis systems, distortion at boundaries [8], artefacts introduced by vocoders [11], F0 statistics [12] exploit knowledge about spoofing algorithms. However, these may not be suited to practical scenarios where the specific spoofing method is not likely to be known in advance.

Recent works have investigated and highlighted the necessity for generalised countermeasures, that do not require prior knowledge of spoofing methods [6]. There were many interesting approaches reported in ASV spoofing challenge 2015 [13], which particularly focused on generalised countermeasure implementation. However, even though most of the systems submitted to this challenge successfully achieved good performance for known attacks, which were seen during enrolment process, many of them failed for unseen attacks, with higher error rates [14]. This suggests that the development of generalised countermeasure still has a long way to go.

The development of features that are able to discriminate between genuine and spoofed speech is an active area of research. For e.g., both SS and VC attacks rely on vocoders and consequently phase based features were shown to provide good discrimination between spoofed and genuine speech [15, 16]. Also, long-term (high-level) features show promise in detecting unseen spoofing attacks [17]. However, to date there have been no investigation of how discriminative information is distributed across frequency bands for spoofing detection.

The use of mel frequency scale filters along the speech bandwidth is the de-facto approach to front-end filter design in most speaker recognition systems. However, it may not be the optimal scale for all speech based task, as previously shown in

http://dx.doi.org/10.21437/Interspeech.2016-844

both speaker verification and cognitive load estimation [18, 19].

The primary aim of this work is to determine how information relevant to spoofing detection is distributed across the various sub-bands of speech. Further, feature extraction approaches that exploit this information are investigated in this work.

2. Database and Performance Metrics

The Spoofing and Anti-Spoofing (SAS) corpus [20] is used in our investigations. This corpus is one of the most diverse and large scale database available for VC and SS based spoofing experiments to date. It consists of three non-overlapped data partitions: train, development and test, including 5 speech synthesis algorithms (2 for training and 3 for testing) and 8 voice conversion algorithms (3 for training and 5 for testing). The training set includes 3750 genuine and 12625 spoofed utterances. The development set comprises 3497 genuine and 152215 spoofed recordings. In the test set there are 9404 genuine and 1034397 spoofed trails. The test set includes additional 8 spoofing attacks of unknown origin, not employed during countermeasure optimization in order to provide a generalized countermeasure evaluation. Only the voiced part of the utterance is used in our experiments, and a voice activity detector used in [21], is employed to remove unvoiced regions.

The primary performance metrics reported in this paper are Equal Error Rate (EER) and Kullback-Leibler (KL) divergence. False Alarm Rate (FAR), which is the ratio between number of spoof trials classified as genuine speech and total number of spoof trials, and Miss Rate (MR), ratio between number of genuine trials classified as spoofed speech and total number of genuine trials, measures are used to derive EER, which are defined as same as the ASVspoof 2015 challenge [22]. By varying the threshold, the trade-of between FAR and MR is determined, i.e. the EER, and reported. KL divergence is generally used to measure the distance between two probabilistic models. We used Monte Carlo approximation based KL divergence [23] to measure the distance between spoofed and genuine speech models.

3. Sub-band Analysis

Initially, an analysis of which sub-bands contain more discriminative information relevant for spoofing detection was conducted. The discriminative ability of different frequency region was estimated using two approaches: model-level comparison and classification-level comparison. The KL divergence between statistical models of genuine and spoofed speech, corresponding to different sub-bands, is used as a model-level measure of discriminative ability. A classification level measure of discriminative ability is also estimated using equal error rate (EER) of a sub-band based spoofing detection system.

3.1. System Description and Configuration

The first phase, for analysing both model-level and classification-level measures of discriminative ability, is to extract features for each sub band as shown in Figure 1. For each frame of speech, frequency bins are divided into sub-bands based on DFT bin groupings. Within each sub-band, DCT is applied to the corresponding log magnitudes to obtain the sub-band features. Each sub-band based measure uses 30 dimensional frame based feature vectors, $\boldsymbol{v_n}$, comprising of 10 DCT coefficients along with the deltas and delta-deltas.

The measure of discriminative ability within each sub-band is determined using a 512 component GMM-UBM system. Within the n^{th} sub-band, the sub-band features, $\boldsymbol{v_n}$, were used to train a UBM following which genuine and spoofed GMM models were obtained via MAP adaptation (mean only) using HTK [24].

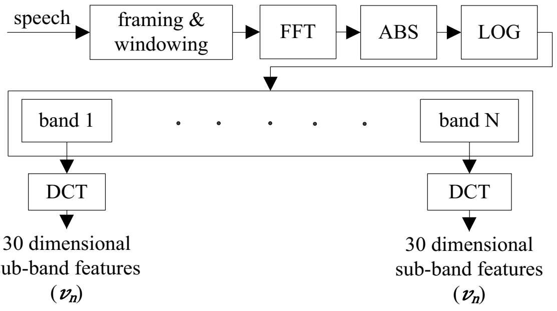

Figure 1: *Block diagram of sub-band based feature extraction.*

Similar to the work of Thiruvaran et al. [18], in the field of speaker verification, the analysis of discriminative information distribution was initially performed across four sub-bands: 0-1kHz, 1-2.5kHz, 2.5-5.5kHz and 5.5-8kHz. Following this, a second approach involved dividing the speech bandwidth into uniform 1 kHz wide sub-bands and the model and classification level analyses were repeated. The two approaches are referred to as 4-band and 8-band divisions in the rest of the paper.

3.2. Model-level Measure

The KL divergences between genuine and spoofed models were estimated within each sub-band and shown in Figure 2. It is evident that some sub-bands have more discriminative information than others (indicated by 'shaded' regions in Figure 2). Specifically, the 0-1kHz, 2.5–5.5kHz and 7-8kHz sub-bands are identified as the most discriminative frequency regions.

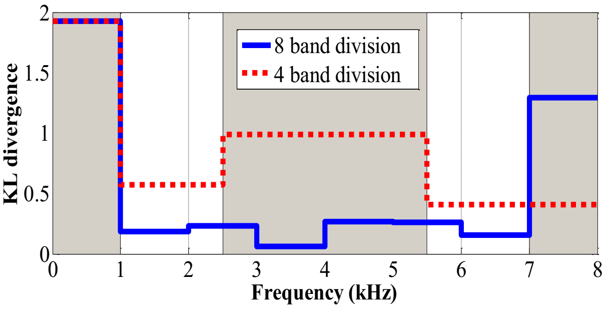

Figure 2: *Comparison of frequency dependency of discrimination between genuine and spoofed speech at model-level. Shaded frequency regions are selected as most discriminative sub-bands.*

3.3. Classification-level Measure

For the classification-level comparison, the Log likelihood (LL) for genuine and spoofed models for each utterance was calculated in order to compute the log likelihood ratio (LLR).

Genuine vs spoofed decision was then made based on these LLRs and the equal error rate (EER) of these decisions was used as the classification-level measure of the discrimination ability within each sub-band. The EERs were estimated for the test trails of the SAS corpus using the MSR toolkit [25]. Figure 3 shows the EER for each sub-band under consideration. It is worth noting that the classification-level EER results agree with the model-level KL divergence results.

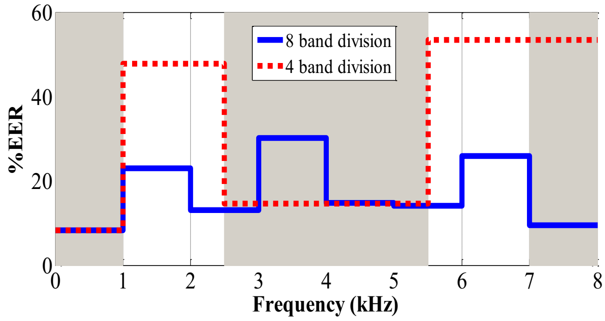

Figure 3: *Comparison of frequency dependency of discrimination between genuine and spoofed speech at classification-level. Low %EER signifies better discriminability and shaded frequency regions are selected as most discriminative sub-bands.*

From Figures 2 and 3, it can be seen that the 0-1 kHz sub-band and the 2.5-5.5 kHz sub-band contain a high proportion of discriminative information that can distinguish between spoofed and genuine speech. It is worth noting that these two sub-bands also contain the most discriminative information for the traditional speaker verification task [18]. In Figure 2, the sub-band 2.5-5.5kHz was identified as one of the most discriminative sub-band, even though the discriminative power in this sub-band is low according to 8-band division approach; however, according to 4-band division approach this sub-band contains more discriminative information. Finally, the 7-8kHz sub-band also carries discriminative information (according to 8-band division approach), which may reflect the nature of most SS and VC algorithms that do not model this frequency region as observed by [26].

4. Proposed Filter Bank Design

As seen in Section 3, the frequency-dependency of discrimination between spoofed and genuine speech does not follow the frequency-dependent density of the mel scale, and hence we propose a new filter design better adjusted to the genuine/spoofed speech separation problem. Sub-band analysis experiment indicated that the 0-1 kHz, 2.5-5.5 kHz and 7-8 kHz sub-bands should be utilised more in feature extraction process, to detect spoofing attacks effectively, which is the motivation for the new filter bank design. The basic idea behind the proposed approaches is the allocation of a greater number of filters within the discriminative sub-bands.

Three different approaches to filter bank design are presented. All three approaches involve assigning the centre frequencies of triangular filters across the speech bandwidth. The first approach is based on simulated annealing to optimise centre frequency positions of the triangular filters and does not make explicit use of the discriminative regions identified in section 3. This method automatically assigns filter banks, through a learning process. The second approach involves the use of only the discriminative frequency regions identified in section 3 and a number of combinations of linearly spaced and mel spaced filters within these regions. The third approach makes use of the KL divergence between models of genuine and spoofed speech within each frequency region to determine the number of filters in those regions. In all three approaches, log energy of the filter outputs were calculated followed by DCT to obtain features (13 DCT coefficients + Δ + ΔΔ) which were used in a two class GMM-UBM structure to distinguish between genuine and spoofed speech as shown in Figure 4. The Gaussian mixture models of both classes were obtained via full MAP adaptation from the UBM.

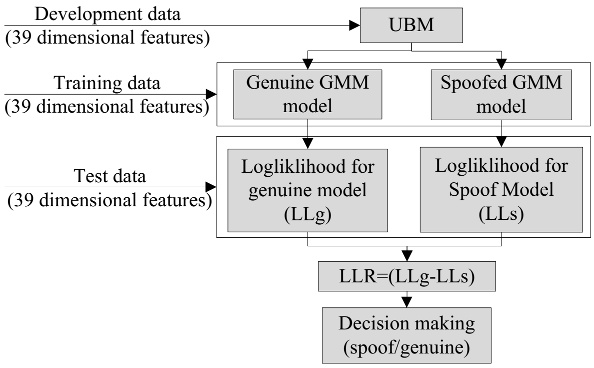
Development data
(39 dimensional features)
UBM
Training data
(39 dimensional features)
Genuine GMM model
Spoofed GMM model
Test data
(39 dimensional features)
Logliklihood for genuine model (LLg)
Logliklihood for Spoof Model (LLs)
LLR=(LLg-LLs)
Decision making (spoof/genuine)

Figure 4: *GMM-UBM structure for genuine vs. spoofed speech classification.*

4.1. Simulated Annealing based Filter Bank

Simulated annealing is used to determine system parameters that maximise (or minimise) a suitable objective function. In this case, the system parameters are the centre frequencies ($\boldsymbol{f_c}$) of the triangular filters and the objective function, $\Phi(\boldsymbol{f_c})$, that is maximised is given by:

$$\Phi(\boldsymbol{f_c}) = D_{KL}\left(\mathcal{G}_g(\boldsymbol{f_c}), \mathcal{G}_s(\boldsymbol{f_c})\right) \tag{1}$$

Where, $D_{KL}\left(\mathcal{G}_g(\boldsymbol{f_c}), \mathcal{G}_s(\boldsymbol{f_c})\right)$ denotes the KL divergence between a Gaussian mixture model of features extracted from genuine speech, $\mathcal{G}_g$, and a Gaussian mixture model of features extracted from spoofed speech, $\mathcal{G}_s$. Since the GMMs are models of the feature space it should be noted that at each iteration of the simulated annealing, the features have to re-estimated with new centre frequencies and the GMMs have to be retrained.

The simulated annealing based filter bank was designed using the MATLAB global optimization toolbox [24]. The annealing temperature was initially set as 100 and an 'exponential temperature update' schedule was followed. The initial values of the centre frequencies were 26 mel-spaced frequencies and the stopping criterion was an average change in the objective function less than a threshold of e^{-2}. The simulated annealing was carried out on 10% of the training set chosen to contain examples of all types of spoofing attacks covered by the entire training set.

Figure 5 shows the filter bank produced by the simulated annealing approach. It can be seen that this method automatically identifies frequency regions with high levels of discriminability between genuine and spoofed speech and allocated more filters in these regions. The frequency regions identified here (Figure 5) tallies with the analyses reported in section 3.

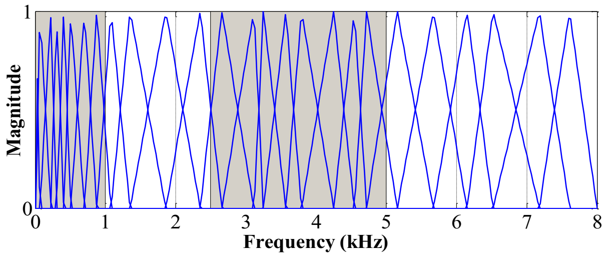

Figure 5: *Filter bank obtained using the simulated annealing approach.*

4.2. Discriminatory Sub-band based Filter Bank

This method focuses only on the sub-bands identified in section 3 as containing discriminative information between genuine and spoofed speech. Four different combinations of linear, mel and inverse-mel (inverting the mel scale from high frequency to low frequency) spaced triangular filters within these sub-bands were implemented and all four corresponding front-ends were compared. In all four front-ends the total number of filters was set as 26 and the number of filters in each sub-band was chosen based on the width of the sub-bands and/or the level of discriminability as determined in section 3.2. The details of the four filter banks are listed in Table 1.

4.3. KL Divergence based Filter Bank

The third approach, to allocate the bandwidths (and consequently centre frequencies) of the triangular filters in each frequency region according to the amount of discriminative information, was implemented as per [19]. The underlying idea behind this filter allocation is to have more filters in frequency regions corresponding to high KL divergence between models of spoofed and genuine speech. The bandwidth of the i^{th} filter (BW_i) is given by:

$$BW_i \approx \frac{\frac{1}{\left(KL(f_i)\right)^{\propto}}}{\sum_{i=1}^{N} \frac{1}{\left(KL(f_i)\right)^{\propto}}} \qquad (3)$$

where, N is the total number of filters, α is a scaling parameter greater than zero, and $KL(f_i)$ is the KL divergence at the centre frequency, f_i, of the i^{th} filter.

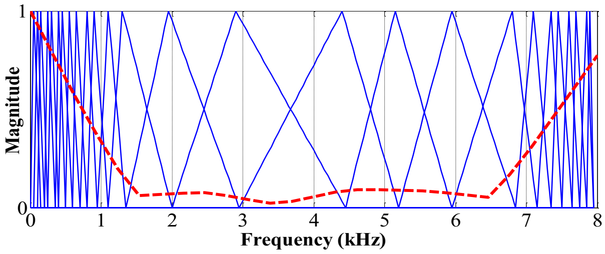

Figure 6: *KL Divergence based Filter Bank obtained with α=1.1. 'Red curve' shows the KL divergence scaled by α as a function of frequency.*

The value of α is empirically chosen and in this case was assigned a value of 1.1. The KL divergence values at different frequencies are obtained from the uniform sub-band analysis (8-band division) described in section 3.2. The final filter bank obtained via this method is shown in Figure 6. It can be seen that there are a greater number of filters in the 0-1 kHz and the 7-8 kHz regions reflecting the correspondingly high values of KL divergence in these regions.

5. Experimental Results

The 2-class GMM-UBM system (Figure 4) was implemented with 512 mixture GMMs and all systems are compared to a baseline 26-filter MFCC based system. Consequently the number of filters in all three proposed approaches was also kept at 26. Performances of the baseline and proposed filter banks are tabulated in Table 1. All the results are evaluated on 'Test' set of SAS corpus in terms of the equal error rate (%EER). The proposed filter banks based on all three approaches outperformed conventional mel and linear filter banks.

Table 1: *Comparison of proposed and baseline filter bank designs for genuine/spoofed speech classification on SAS corpus.*

	Filter banks	**%EER**
Baseline	26-Linear filter bank	8.91
	26-Mel filter bank (MFCC)	6.89
Simulated Annealing	Filter bank using SA	6.28
Discriminatory Sub-band	10Linear filter in 0-1kHz + 16Linear filter in 2.5-5.5kHz	6.57
	17Linear filter in 0-1kHz + 9Linear filter in 2.5-5.5kHz	6.54
	13Mel filter in 0-1kHz + 13 Inverse-Mel filter in 7-8kHz	6.14
	11Mel filter in 0-1kHz + 7Linear filter in 2.5-5.5kHz + 8 Inverse-Mel filter in 7-8kHz	6.09
KL Divergence	Filter bank according to KL divergence (Section 4.1.3)	6.39

6. Conclusions

In this paper, we have used two different measures (at a model-level and at a classification-level) to identify sub-bands that contain discriminative information between genuine and spoofed speech in the context of automatic speaker verification. Three such discriminatory sub-bands were identified: 0-1 kHz, 2.5-5.5 kHz and 7-8 kHz. We have then proposed three approaches to design banks of triangular filters that allocate a greater number of filters to the more discriminative sub-bands, where simulated annealing and KL divergence based approaches assign the number of filters automatically based on discriminative information between genuine and spoofed speech. All three approaches were experimentally validated on the SAS corpus and outperform conventional mel-spaced and linearly-spaced filter banks. The number of filters in the filter bank is a key parameter that could have a significant impact on system performance. In this work, 26 filters were chosen to match the baseline system, but the optimising the number of filters in each sub-band is an avenue for future work.

7. Acknowledgements

The authors would like to thank Dr. Zhizheng Wu, Centre for Speech Technology Research, University of Edinburgh, for providing the SAS corpus used in the experiments reported in this paper.

8. References

[1] 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 *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, 2012, pp. 1-5.

[2] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4401-4404.

[3] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*, 2004, pp. 145-148.

[4] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Biometrics and ID Management*, ed: Springer, 2011, pp. 274-285.

[5] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 20, pp. 2280-2290, 2012.

[6] 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.

[7] E. Khoury, T. Kinnunen, A. Sizov, Z. Wu, and S. Marcel, "Introducing i-vectors for joint anti-spoofing and speaker verification," in *Proceedings interspeech*, 2014.

[8] F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 3068-3072.

[9] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "The AHOLAB RPS SSD Spoofing Challenge 2015 Submission," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] L.-W. Chen, W. Guo, and L.-R. Dai, "Speaker verification against synthetic speech," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, 2010, pp. 309-312.

[11] Z. Wu, C. E. Siong, and H. Li, "Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition," in *INTERSPEECH*, 2012, pp. 1700-1703.

[12] A. Ogihara, U. Hitoshi, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE transactions on fundamentals of electronics, communications and computer sciences,* vol. 88, pp. 280-286, 2005.

[13] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah*, et al.*, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," *Training,* vol. 10, p. 3750, 2015.

[14] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "Spoofing and Anti-Spoofing: A Shared View of Speaker Verification, Speech Synthesis and Voice Conversion," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Hong Kong, China, 2015.

[15] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *Information Forensics and Security, IEEE Transactions on,* vol. 10, pp. 810-820, 2015.

[16] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[17] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, 2013, pp. 1-8.

[18] T. Thiruvaran, V. Sethu, E. Ambikairajah, and H. Li, "Spectral shifting of speaker-specific information for narrow band telephonic speaker recognition," *Electronics Letters,* vol. 51, pp. 2149-2151, 2015.

[19] P. N. Le, E. Ambikairajah, E. H. Choi, and J. Epps, "A non-uniform subband approach to speech-based cognitive load classification," in *Information, Communications and Signal Processing, 2009. ICICS 2009. 7th International Conference on*, 2009, pp. 1-5.

[20] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda*, et al.*, "SAS: A speaker verification spoofing database containing diverse attacks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 4440-4444.

[21] M. Nosratighods, T. Thiruvaran, J. Epps, E. Ambikairajah, M. Bin, and L. Haizhou, "Evaluation of a fused FM and cepstral-based speaker recognition system on the NIST 2008 SRE," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 4233-4236.

[22] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training,* vol. 10, p. 3750, 2014.

[23] V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in speech based emotion models-Analysis and normalisation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7522-7526.

[24] M. A. Branch and A. Grace, *MATLAB: optimization toolbox: user's guide version 1.5*: The MathWorks, 1996.

[25] 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,* 2013.

[26] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.