PHONEME SPECIFIC MODELLING AND SCORING TECHNIQUES FOR ANTI SPOOFING SYSTEM

Gajan Suthokumar [1,2], Kaavya Sriskandaraja [1], Vidhyasaharan Sethu [1], Chamith Wijenayake [1], Eliathamby Ambikairajah [1,2]

[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2]DATA61, CSIRO, Sydney, Australia

g.suthokumar@unsw.edu.au, k.sriskandaraja@unsw.edu.au, v.sethu@unsw.edu.au, c.wijenayake@unsw.edu.au, e.ambikairajah@unsw.edu.au

ABSTRACT

Replay attack refers to the use of recorded speech in an attempt to spoof an automatic speaker verification system and the development of countermeasures that can detect these attacks is an active area of research. This paper investigates the effect of phoneme specific information on replay attack detection. It then develops a replay detection system that employs phoneme specific genuine and spoof models and compares novel scoring methods that take into account phonetic information obtained from a suitable phoneme recogniser. Experiment result on the ASVSpoof 2017 V2.0 corpus indicated that replayed speech may be easier to detect from speech corresponding to some phonemes compared to others and consequently judicious use of phoneme specific models can improve replay detection systems.

Index Terms— spoofing detection, replay attack, phoneme detection, phoneme posterior weighted score, speaker verification.

1. INTRODUCTION

Replay attacks are simple yet effective means by which automatic speaker verification (ASV) system can be spoofed using simple audio record and playback devices [1]. Most current approaches to replay detection rely on the observation that the speech signal involved in replay attacks must pass through both recording and playback channels, which in turn may result in some spectral distortion. Replay detection may then be cast as a problem of detection of this channel distortion, while taking into consideration that there is a myriad of recording and playback channels and these cannot be known a priori.

Generally, spoofing detection includes front-end as well as back-end and most of the anti-spoofing research for replay attack has been focused on feature engineering while the classification blocks are often built on the traditional classification techniques such as Gaussian mixture model (GMM), support vector machine (SVM) [2]. Front ends based on variants of spectral features, long-term spectral statistics [3], voice source [4], phase based features [5] and different variants of deep neural network based systems [5 - 9] have been investigated, extensively. The features indicative of spectral cues, include spectral centroid magnitude coefficient (SCMC) [10], constant-Q cepstral coefficient (CQCC) [11], rectangular filter cepstral coefficients (RFCC) [10], scattering coefficients [12], spectral energy slope [13], spectro-temporal modulation feature (STMF) [14, 15], often use spectrogram to extract the information. It has been suggested that replayed signals would include noise and reverberation, leading to a flatter and altered spectrogram [14]. Each region of spectrogram tends to be affected differently which in turn could mean different phonemes are affected differently by the replay channel. Furthermore, it has also been suggested that different phonemes vary in their robustness to reverberation in the context of automatic speech recognition [16]. Motivated by above findings, we aim to investigate how phoneme related information can be incorporated in to replay attack detection systems.

In addition, features employed in any spoofing detection system will be incorporate variability due to a number of factors such as channel effects, differences between speakers, and phonetic variability arising from the linguistic content. Previous work has shown that replay detection can be improved by making use of speaker specific models and in turn implicitly compensating for speaker variability [15]. Phonetic variability is generally not explicitly taken into consideration. Instead, most back-ends model the statistical distribution of the features for replayed and genuine speech and rely on the back-ends capturing the differences across all phonemes. However, in other areas of speech processing, such as emotion recognition [17] and speaker verification [18], explicit modelling of phonetic information has been shown to be beneficial.

This paper makes three key contributions; firstly we investigate if some phonemes are more conducive replay detection than others; secondly, we proposed a novel framework to incorporate phoneme specific models into a replay detection system; and finally we compare four scoring methods developed to incorporate phonetic information. To the best of the authors' knowledge this is the first study on the effect of phonetic variation in replay detection.

2. DATABASE

The original ASVSpoof 2017 challenge corpus [19], comprising of genuine recordings and their replayed versions, are used in all the experiments outlined in this paper. The RedDots text dependent corpus is used directly for the genuine utterances. Replayed speech utterances are created through recording the playback of the genuine speech through the different playback and recording devices in various acoustic environments. Three non-overlapping subsets as train, development and evaluation are provided. As this is a text dependent corpus, 10 phrases have been used in all subsets. Anomalies identified in the original ASVSpoof 2017

corpus prompted the organisers to release an updated version referred to as the ASVSpoof 2017 Version 2.0 (V2.0) corpus and a new enhanced CQCC baseline in 2018 [11] and all our experiments results reported here are on the V2.0 corpus . It should be noted that results reported using the original ASVSpoof 2017 (V1.0) are not directly comparable with V2.0 results.

3. PHONETIC VARIABILITY ANALYSIS

As previously mentioned, replay detection may be cast as a problem of detecting an a priori unknown channel, where the channel comprises of the recording and playback devices in addition to the acoustic environment. This in turn is typically implemented as the detection of the spectral distortion introduced by the channel. Consequently, since different phonemes have different spectral characteristics (for instance, fricatives have more of their energy contents in the high frequency regions while vowels contain more of their energy in the lower frequency regions), the ease of detecting the spectral characteristics of an unknown channel may vary across different phonemes.

Our main aim of this work is to determine whether some phonemes allow easier detection of spoofed speech compared to others. Specifically, we investigate whether every phoneme affected differently, during the process of replay. To analyse which phoneme has more discriminative ability each phoneme has examined separately. First, the corresponding phoneme presents in each frame is detected and then discriminative power between genuine and spoof class of different phonemes was estimated using two approaches: model-level and classification-level comparison.

3.1. Phoneme Detection and Frame Labelling

A frame based phoneme detector is used to estimate the phoneme posteriors for each frame. The most dominant phoneme (the phoneme with the highest posterior probability) was then associated with the corresponding frame, determined by applying a predetermined threshold to the phoneme posterior probabilities.

Table 1. *Categories of phoneme*

Vowels		aa, ae, ah, eh, er, ih, iy, uh, uw
Semivowels		l, r, w, y
Diphthongs		aw, ay, ey, ow, oy
Affricatives		ch, jh
Fricatives	**Voiced**	dh, dx, v, z
	Unvoiced	f, hh, th, s, sh
Nasals		m, n, ng
Stops	**Voiced**	b, d, g
	Unvoiced	k, p, t
Non speech/Silence		pau

The BUT phoneme recognizer [20] with 39 English phonemes is used to identify the relevant phoneme label through the phoneme posterior probabilities. When evaluated on the TIMIT database, phone recogniser had an accuracy of 74%. A posterior probability threshold of 0.75 is empirically determined which assigns a reasonable number of frames to each phoneme. This means a frame with a highest phoneme posterior probability of greater than 0.75 is assigned to the corresponding phoneme label or otherwise it will be discarded without any label assignment to ensure that only frames with high likelihood of corresponding to one of the phonemes are chosen for train data. This thresholding retained about 15% of original training data. The phonemes can be classified (as in Table 1) as either vowels, semivowels, diphthongs, affricatives, fricatives, nasals, stops and silence ('Pau'). 'Pau' phonemes (i.e. non-speech) are also included since they tend to carry artefacts of the acoustic environment channels present in replayed speech and they have been shown to be effective in replay detection [21, 22].

3.2. Individual Spoofing Detection

In order to measure the discriminative ability of each phoneme, a spoofing detection system is setup and the discriminative ability measures for different phonemes are determined, independently. For this purpose, as shown in Figure 1, individual genuine and spoof models for every phoneme are utilized for spoofing detection.

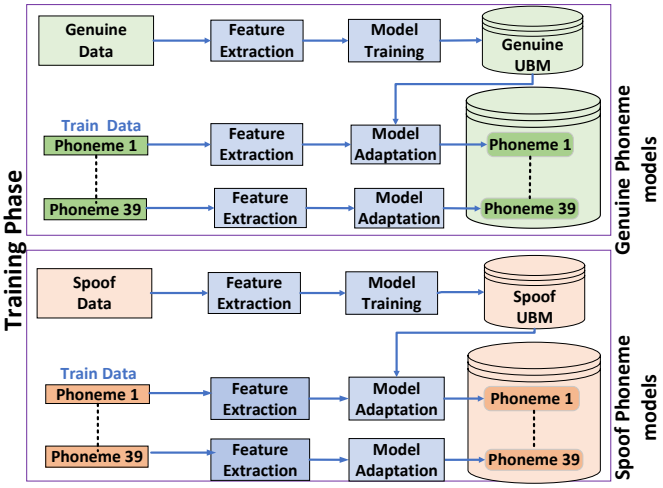

Figure 1: *Training of phoneme specific models*

3.2.1. Feature Extraction

Rectangular frequency cepstral coefficients (RFCCs) are state-of-the-art short-term (frame based) features reported on the ASVspoof V2.0 of the corpus [10, 21], and are used in the systems developed in this work. RFCCs were computed with a 20ms frame duration obtained using a hamming window with 50% overlap. The speech was pre-emphasized prior to feature extraction. RFCCs were extracted using a linear frequency scale as it captures the replay channel artefacts better than other frequency scales, e.g. mel, inverse-mel etc [10]. 40-dimensional RFCCs were extracted and appended with their dynamic coefficients (velocity and acceleration) to obtain a 120-dimensional feature vector. Cepstral mean normalization (CMN) is then carried, which was found to be highly beneficial for replay detection [10]. Previously proposed state-of-the-art spectro temporal modulation feature (STMF) (utterance level feature) [14] and other longer term features are not suitable for use in this work as the frame duration must be less than typical phoneme durations.

3.2.2. Modelling and Classification

The approach proposed in this work to study phoneme variability employs GMM based models of spoofed and genuine speech. Specifically, a background model for genuine and spoofed speech is initially trained on genuine and spoofed speech, using the EM

algorithm, and is referred to as the genuine universal background model (genuine UBM) and spoofed UBM. Following this, each phoneme models are adapted from genuine UBM and spoofed UBM (refer Figure 1).

Given these models and a frame of a test utterance, the final decision about whether it is genuine speech or replayed speech is based on the log-likelihood ratio (LLR) between genuine and spoofed model of corresponding phoneme is given by:

$$LLR_j^{(i)}(X_j) = log\ P\left(X_j \middle| \theta_g^{(i)}\right) - log\ P\left(X_j \middle| \theta_s^{(i)}\right) \quad (1)$$

where X_j denotes the feature vector from j^{th} frame of the test utterance, $\theta_g^{(i)}$ denotes the genuine GMM corresponding to i^{th} phoneme, $\theta_s^{(i)}$ denotes the spoofed GMM belongs to i^{th} phoneme. Rather than assigning a frame to a phoneme we have used the phoneme posteriors as weight to each LLR score to make the final decision as follows:

$$weighted\ LLR_j^{(i)}(X_j) = p_j^{(i)}(X_j) \times LLR_j^{(i)}(X_j) \quad (2)$$

where $p_j^{(i)}$ is the phoneme posterior probability of corresponding frame which calculated from the BUT phoneme detector in our case. The reasons behind weighting LLR by phone posterior probability is to retain all the frames in the test utterance while incorporating the information about how likely each frame belongs to the 39 phonemes. This procedure differs from training phase where the frames with at least 0.75 probability are chosen for training.

A 512 mixture GMM is used for each of the genuine and spoof UBM models, which are estimated using the Expectation Maximization (EM) criteria on the pooled train and development sets. Phoneme specific models are learnt via mean and variance MAP adaptation from corresponding UBMs using train data (phoneme labelled data after the thresholding as described in Section 3.1) with MAP adaption factor of 10 which is empirically chosen from development set.

3.3. Baseline System

The current replay detection systems employ a 'genuine' speech model and a 'spoofed' speech model that is modelled using genuine and spoofed training data. The log likelihood ratio of a frame (LLR_j) and final score (LLR) for an utterance for the baseline system are generally calculated as shown below:

$$LLR_j\ (X_j) = log\ P(X_j|\theta_g\ ) - log\ P(X_j|\theta_s\ ) \quad (3)$$

$$LLR(X) = \frac{1}{N}\sum_{j=1}^{N} LLR_j\ (X_j) \quad (4)$$

where X_j denotes the set of feature vector from j^{th} frame of the test utterance X, θ_g denotes the genuine GMM and θ_s denotes the spoofed GMM and N denotes the total number of frames of the utterance. The number of mixture components for the GMM used in the baseline was empirically chosen to be 512, with no performance increment observed when using more components.

3.4. Measures of Discriminability

In order to discern the effect of the different phoneme, we carried out the model level and classification level discriminative analysis between the genuine and spoofed speech using the Kullback-Leibler (KL) divergence and Equal Error Rate (EER), respectively for every phoneme independently. 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 phoneme specific spoofed and genuine speech models. False alarm rate (FAR) and miss rate (MR) measures are used to derive EER, which are defined as same as the ASVspoof 2017 challenge [19]. By varying the threshold, the trade-off between FAR and MR is determined. The EER is a single point where trade-off leads to equal value for FAR and MR.

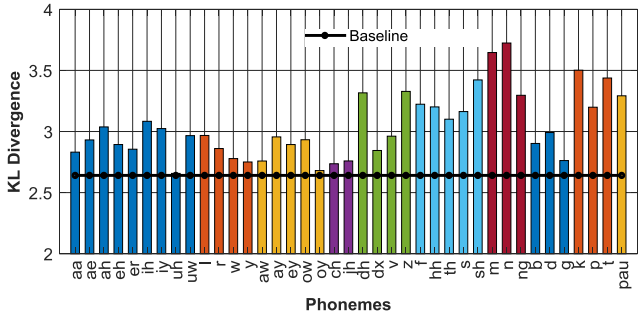

Figure 2: *KL divergence of the 39 phonemes estimated between their respective genuine and spoof GMMs with baseline system on ASVSpoof 2017 V2.0 train set. Here X-axis (phonemes) follows the phoneme category order tabulated in Table 1.*

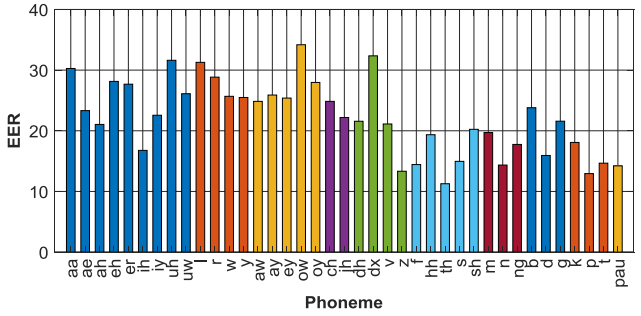

Figure 3: *Comparison of %EER of individual spoofing detection system of 39 phonemes on ASVSpoof 2017 V2.0 development set. Here X-axis (phonemes) follows the phoneme category order tabulated in Table 1.*

KL divergence of the 39 phonemes are estimated between their respective genuine and spoof GMMs in the ASVSpoof 2017 V2.0 train set independently. KL divergence between the genuine and spoof models for each phoneme and baseline genuine and spoof model is shown in Figure 2. A larger degree of mutual dissimilarity results a large KL divergence and vice versa. It is clearly noticeable that the KL divergence of individual phoneme models varies among them and all of them have the higher KL divergence than the baseline system. It is evident that phoneme specific modelling of genuine and spoofed speech could be more beneficial than the modelling one phoneme independent genuine and spoof models each.

Classification level measure is calculated in terms of EER of each phoneme on ASVspoof 2017 V2.0 development set independently, using the weighted LLR as described in the equation 2 and depicted in Figure 3. Lower EER describes the high discrimination. It is evident that the EER of every phoneme differ with each other as similar to the findings from KL divergence plot. Also, it is noticeable that the phonemes correspond to the higher

EER ones have lower KL divergence and lower EER ones have higher KL divergence and vice versa.

From Figure 2 and Figure 3, it can be seen that every phoneme has different levels of discriminative ability. It is clear that almost all the fricatives, nasals, stops and silence regions have been helping more towards identifying the replay channels than most of vowels, semivowels, diphthongs and affricatives. This is arguably true since vowels have more energy in low frequencies and while fricatives have more energy in high frequencies as replayed speech have been shown to exhibit more artefacts at higher frequencies [8]. Also, within fricatives and stops it is been observed that the unvoiced ones are more discriminative than the voiced ones. This analysis also confirmed that the Silence ('Pau') regions are helpful as they have the emphasized environmental channel artefacts than speech regions, which are also masked with the speaker related information.

4. PROPOSED PHONEME SPECIFIC SYSTEM AND SCORING TECHNIQUES

As seen in Section 3, the phoneme-dependency of discrimination between genuine and spoofed speech varies for each and every phoneme, and hence we propose a novel framework to incorporate phoneme specific models into a replay detection system and four scoring methods developed to incorporate phonetic information. Phoneme specific spoofing detection system consists of genuine and spoofed models for each individual phoneme independently (as shown in Figure 1) and score level fusion.

Phonetic variability analysis experiment indicated that some phonemes which carry more information should be utilized more in decision making process, to detect replay attack effectively, which is in the motivation for the new scoring techniques. There are four scoring methods proposed in this paper. The first two techniques simply make use of scores derived from each phoneme models for the fusion. These are referred to as phone posterior weighted score (PPWS) and maximum phone posterior weighted score (PPWS_max) and defined as:

$$PPWS = \frac{1}{N}\sum_{i=1}^{39}\sum_{j=1}^{N} p_j^{(i)}(X_j) \times LLR_j^{(i)}(X_j) \quad (5)$$

$$PPWS\_max = \frac{1}{N}\sum_{j,i=Max(p_j^{(i)})}^{N} p_j^{(i)}(X_j) \times LLR_j^{(i)}(X_j) \quad (6)$$

where N denotes the total number of frames of the utterance. $p_j^{(i)}$ and $LLR_j^{(i)}$ are the phoneme posterior probability and LLR of the j^{th} frame for the corresponding i^{th} phoneme model which is defined as, $p_j^{(i)} = P(Phone_i|X_j)$ with $\sum_{i=1}^{39} P(Phone_i|X_j) = 1$. The knowledge of the $Phone_i$ can be extracted using phoneme recognizer.

The other two techniques assign a weight for each separate phoneme clusters explicitly which will be referred as phoneme posterior and relevance factor weighted score (PPRFWS) as shown below:

$$PPRFWS = \frac{1}{N}\sum_{i=1}^{39} C_i \sum_{j=1}^{N} p_j^{(i)}(X_j) \times LLR_j^{(i)}(X_j) \quad (7)$$

where C_i is phoneme relevance factor for i^{th} phoneme.

The phoneme with high discriminability will have a higher C_i which is estimated in two ways. One is based on explicit assumption about the data which is derived using KL divergence on the development set. The KL based PPRFWS is computed by taking the $C_i = \frac{KL_i}{\sum_1^{39} KL_i}$ which will be referred as PPRFWS_KL. The other is completely learnt on some dataset with not specific assumptions where the phoneme relevance factors $C_i, i \in [1,39]$ is learned by the linear regression (LR) classifier where the objective function targets to minimize the EER on the development set, which will be referred as PPRFWS_LR.

5. EXPERIMENTAL RESULTS

The proposed scoring methods are used to fuse the scores from 39 phoneme specific models (refer Figure 1) and the results are tabulated in Table 2. These results show that PPWS scoring, which treats all the phoneme classes equally, is outperformed by PPRFWS_KL and PPRFWS_LR, which involve class-wise weightings. Additionally, PPWS_max which assigns a frame to a single phoneme does not perform as well any of the other proposed scoring methods that fuse all 39 scores.

Additionally, while state-of-the-art features such as STMF features, which give an EER of 7.2% on ASVspoof 2017 V2.0 evaluation set, cannot be employed in the proposed framework since they are long term (utterance level) features, but can be fused with the proposed systems. This was evaluated by fusing the proposed PPRFWS_LR system with the STMF system [14] which resulted in an EER of 6.18%. Finally, we carried out some empirical tests to determine if dropping scores corresponding to any of the phonemes provided any benefit but it was observed that the best results were always obtained when all 39 scores were taken into consideration. Also, similar finding is observed for SCMC [10, 21] features which is the second best state-of-the-art short term frame level feature.

Table 2. *Evaluation results in terms %EER for the baseline system and the proposed systems on ASVSpoof2017 V2.0 corpus (Pooled train and development set is used for training phase)*

	System	%EER
Baseline	RFCC [10, 21]	11.22
Proposed Systems	PPWS	10.70
	PPWS_max	11.57
	PPRFWS_KL	9.97
	PPRFWS_LR	**9.28**
Fusion	STMF [14] + PPRFWS_LR	6.18

6. CONCLUSION

This work investigates how the ability to discriminate between genuine and spoofed speech varies across different phonemes and the consistently higher level of discriminability for frames associated with certain phonemes, especially fricatives, nasals, stops and pause indicates that these types of phonemes are more informative in the detection of replay attacks. We have then proposed four different fusion scoring methods to incorporate phonetic information using phoneme specific models of genuine and spoofed speech and experimental validation on the ASVspoof2017 V2.0 corpus demonstrates that all approaches that take into account all the phoneme specific models outperform the baseline phoneme independent modelling approach.

7. REFERENCES

[1] Z. Wu, S. Gao, E. S. Cling, H. Li, E. S. Chng, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," *2014 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA 2014*, 2014.

[2] L. Li, Y. Chen, D. Wang, T. Fang Zheng, and T. F. Zheng, "A Study on Replay Attack and Anti-Spoofing for Automatic Speaker Verification," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, vol. 2017–Augus, pp. 92–96.

[3] H. Muckenhirn, P. Korshunov, M. Magimai-Doss, and S. Marcel, "Long-Term Spectral Statistics for Voice Presentation Attack Detection," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 11, pp. 2098–2111, 2017.

[4] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features," in *Interspeech*, 2017.

[5] F. Tom, M. Jain, P. Dey, and I. Kharagpur, "End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention."

[6] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Interspeech*, 2017, pp. 82–86.

[7] P. Shih, C. Chen, and H. Wang, "Speech emotion recognition with skew-robust neural networks," in *Icassp*, 2017, pp. 2751–2755.

[8] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination," in *Interspeech*, 2017, pp. 97–101.

[9] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric," in *Interspeech*, 2018, pp. 671–675.

[10] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-Results on the ASVspoof 2017 Challenge," in *Interspeech*, 2017, pp. 7–11.

[11] M. Todisco, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," in *Odyssey*, 2018, pp. 296–303.

[12] K. Sriskandaraja, G. Suthokumar, V. Sethu, and E. Ambikairajah, "Investigating the use of scattering coefficients for replay attack detection," in *In Proceedings, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1195–1198.

[13] S. M. S and H. A. Murthy, "Decision-level feature switching as a paradigm for replay attack detection," in *Interspeech*, 2018, pp. 686–690.

[14] G. Suthokumar, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Modulation Dynamic Features for the Detection of Replay Attacks," in *Interspeech*, 2018, pp. 691–695.

[15] G. Suthokumar, K. Sriskandaraja, V. Sethu, C. Wijenayake, E. Ambikairajah, and H. Li, "Use of Claimed Speaker Models for Replay Detection," in *In Proceedings, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018.

[16] P. P. Parada, D. Sharma, P. A. Naylor, and T. van Waterschoot, "Reverberant speech recognition: A phoneme analysis," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 567–571.

[17] V. Sethu, E. Ambikairajah, and J. Epps, "Phonetic and speaker variations in automatic emotion classification," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 617–620, 2008.

[18] J. Ma, V. Sethu, E. Ambikairajah, and K. A. Lee, "Speaker-Phonetic Vector Estimation for Short Duration Speaker Verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5264–5268.

[19] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Interspeech*, 2017, pp. 2–6.

[20] P. Schwarz, P. Matejka, and J. Cernocky, "Heirarchical Structures of Neural networks for phoneme recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2006, pp. 325–328.

[21] G. Suthokumar, K. Sriskandaraja, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Independent Modelling of Long and Short Term Speech Information for Replay Detection," in *Speech Science and Technology Conference (SST)*, 2018.

[22] MS Saranya, R. Padmanabhan, and H. Murthy, "Replay attack detection in speaker verification using non-voiced segments and decision level feature switching," in *SPCOM*, 2018, pp. 332–336.

[23] V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in speech based emotion models - Analysis and normalisation," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 7522–7525, 2013.