Replay Attack Detection in Speaker Verification Using non-voiced segments and Decision Level Feature Switching

Saranya M. S.[1], Padmanabhan R.[2], Hema A. Murthy[1]
[1] Department of CSE, Indian Institute of Technology Madras, India
[2] School of Computing & Electrical Engg., Indian Institute of Technology Mandi, India
Email: [1]{saranms, hema}@cse.iitm.ac.in, [2]{padman@iitmandi.ac.in}

***Abstract*— This paper proposes a novel approach for replay attack detection, using reverberation and channel information from non-voiced (silence and unvoiced) segments of utterances. The non-voiced segments are determined using a voice activity detector. These non-voiced segments are likely to contain reverberation and channel information. Multiple feature representations are used to capture the remnant vocal tract information in non-voiced segments. Gaussian mixture models are used to build three different baseline systems corresponding to that of three different features. Voting is performed to decide whether a given input utterance is replayed or not. Equal error rate (EER) is computed using the likelihood ratio of the genuine and spoofed model from the best baseline system. Evaluation on the ASV-Spoof-2017 challenge dataset shows that the proposed approach outperforms the best baseline system with a relative improvement of 37% in terms of EER.**

I. Introduction

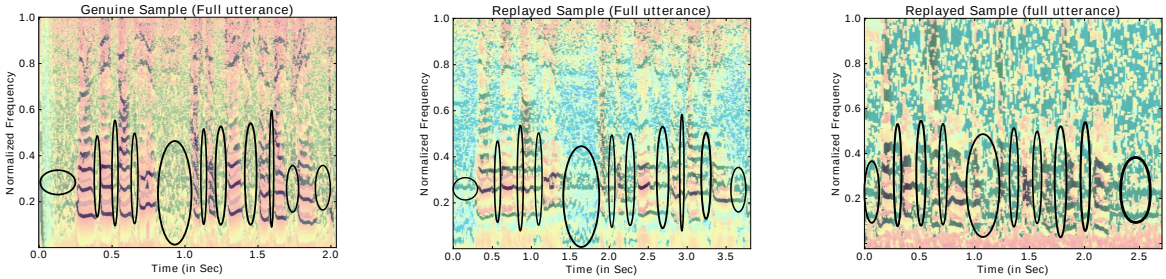
Spoofing is a process, where an impostor gains illegitimate access to an automatic speaker verification (ASV) system, by masquerading as an authentic speaker [1]. One form of spoofing is a replay attack, in which the pre-recorded authentic utterance of a genuine speaker is used by an impostor to access the ASV system. Owing to the small form factor of smartphones and inexpensive, high-quality recording and playback devices, replay attacks have begun to challenge ASV systems. Hence effective detection of such attacks is crucial in securing ASV systems.

ASV-Spoof 2017 database has been designed for detection of replay attacks [2]. Various features were used in the recent literature for the replay detection task. Linear Frequency Cepstral Coefficients (LFCC), and Inverse Mel-Frequency Cepstral Coefficients (IMFCC) were used in [3]. Sub-band analysis on IMFCC and LPCC residual features were performed in [4]. Single Frequency Filtering Cepstral Coefficients (SFFCC) and High-frequency cepstral coefficients (HFCC) were proposed in [5] and [6] respectively. [7] gives a comprehensive study of the relevance of nine different features for replay utterance detection. Authors in [8] used Instantaneous Frequency (IF) with Variable length Teager Energy Operator (VTEO) based Energy Separation Algorithm (ESA) to identify the replayed utterances. A replay detection system trained using the features extracted from a Light Convolutional Neural Network (LCNN) with Max-Feat-Map (MFM) activation function [9] gives best performance on the evaluation data. All these works use MFCC and CQCC features as baseline features, and a Gaussian Mixture Model (GMM) classifier or a Deep Neural Networks (DNN) or a Support Vector Machine (SVM) for final classification. In all these works, the best performing system is the late fusion system which fuses scores from different feature/classifier based replay detection systems.

A replayed utterance has the information of the recording device (R), replayed device (P) and the recorded environment (E), along with the information in the corresponding genuine utterance [10]. The E-R-P information and the effect of reverberation are the additional channel information in the replayed utterance. Hence effectively utilizing channel information is an important aspect in detecting replayed utterances. In [11], for text-dependent ASV systems, a spectral peak mapping method is proposed as a countermeasure, to detect the replay attack on a remote telephone interaction. In [12], countermeasures for the replay attacks with concatenated recordings were proposed. MFCC and pitch are used together to detect channel and pitch mismatch in the concatenated replayed utterance. Villalba in [13] addressed replay attacks with far-field recordings. In [14] the channel details of authentic recordings are collected without any speech, and a channel model is trained. The test utterances with unknown channel information are declared as replayed utterances. Obtaining prior knowledge of all test utterances' channel information is practically not feasible.

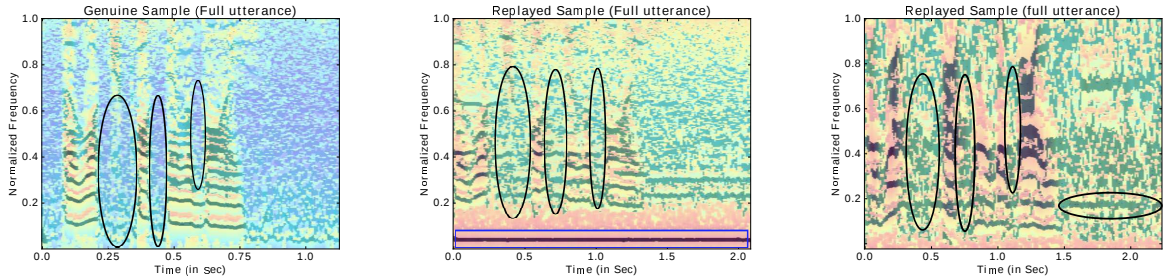
In this paper, instead of collecting silent recordings with expected channel information, Voice Activity Detection (VAD) [15] is used to get the boundaries of silence (S), voiced (V) and unvoiced (Uv) segments of an utterance. The non-voiced segments (S+Uv) of an utterance are expected to contain channel and reverberation information. Thus, in this work, the non-voiced segments obtained using VAD are concatenated and used for the replay detection task. The reverberation and channel information in the replayed utterances vary for different combinations of E-R-P. Hence, a single feature representation may not be adequate to detect different instances of replay attacks. Feature switching refers to the paradigm of utilizing different feature representations for different classes [16], [17]. The novelty of the proposed approach is to use the concept of feature switching at the

978-1-5386-3821-7/18/$31.00 © 2018 IEEE



(a) (b) (c)

Fig. 1: Subplot (a) is the pyknogram of a genuine utterance of a phrase. Subplots (b) and (c) are the pyknograms of the replayed utterances of same phrase captured in office and bedroom environment respectively. The encircled non-voiced regions which have noise-like pattern in genuine utterance and has decaying reverberation tails of formants in the replayed utterances.



(a) (b) (c)

Fig. 2: Subplot (a) is the pyknogram of a genuine utterance of another phrase. Subplots (b) and (c) are the pyknograms of the replayed utterances of same phrase captured in bedroom and office environment respectively. The encircled non-voiced regions have noise-like pattern in genuine utterance and has decaying reverberation tails of formants in the replayed utterances.

decision level, along with the information from the non-voiced segments to detect the replayed utterance.

To facilitate feature switching at the decision level in a Gaussian Mixture Model (GMM) framework, the following strategy is employed: Three different GMM based replay detection systems are built with three features namely: Constant-Q Cepstral Coefficients (CQCC) [18], MFCC, and Mel-Filterbank-Slope (MFS) based features. These three systems are henceforth termed as baseline systems. MFS features proposed in [19], is used for the *first time*, to detect the replay attacks. The log-likelihood ratio score ($\mathcal{S}$) for every test utterance is computed between genuine and impostor models using baseline systems. Based on the baseline scores, every trial is labeled as "genuine/spoofed". For every trial, the baseline scores $\mathcal{S}$ and the voted label are used to make the final decision. Thus for a particular trial, the final score may come from any one of the three baseline systems and in turn, the final score for a set of trials will be from different feature spaces. This approach of making the final decision by choosing the scores from different baseline systems is termed as *Decision Level Feature Switching* (DLFS), which perhaps depends on the environment that is chosen.

The rest of this paper is organized as follows: Section II describes the ASV-Spoof-2017 database. The role of non-voiced segments in detecting the replay attacks is discussed in Section III. The proposed DLFS systems are elaborated in Section IV. The experimental setup is detailed in Section V. The results of the baseline systems and the proposed system are analyzed in Section VI followed by the conclusion in Section VII.

II. Database Description

ASV-Spoof-2017 corpus is a subset of RedDots data collection detailed in [20] and its replayed derivatives. The replayed trials are generated in wild conditions, with different recording and playback environments/devices. The dataset is divided into three subsets namely, training, development, and evaluation. Evaluation data comprises of many utterances collected from new speakers in new E-R-P conditions. The new set of E-R-P conditions are not seen in the development or training data, hence pose a challenge when handling an *unseen* speaker or channel.

III. Reverberation Information

The speaker and speech information like prosody, spectral attributes, and high-level lexical features will be identical between genuine and replayed instances of an utterance while the channel information is likely to be different [21]. A legitimate utterance will have the channel information of the recording device, whereas the replayed utterance will have information about three different channels as mentioned in Section I [10]. Along with the additional channel information, the replayed utterances also have reverberation information.

Reverberation is a decaying tail that occurs when a source is convolved with that of the room response. The reverberation of less than 0.5 seconds is not likely to be perceived by humans, but it is likely to be present in the recorded and replayed signal.

Figure 1(a) is the pyknogram [22] of the phrase "Birthday parties have cupcakes and ice-creams" from ASV-Spoof-2017, spoken by a genuine speaker in an office environment. This instance was recorded with an H6 handy recorder and played back through high-quality studio monitors in bedroom and office environment respectively (Figures 1(b)) and 1(c)). Similarly, Figures2 (a),(b) and (c) are the pyknograms of the genuine and replayed utterances of the phrase "A watched pot never boils" spoken by the same speaker. In this case, the replayed utterance is recorded using the same H6 handy recorder in the bedroom environment and played back through the Creative-A-60 speakers connected to a laptop. The horizontal band highlighted in the Figure 2(c) is due to the low-frequency buzz created by the playback device. The ellipses marked in the figures highlight the decaying reverberation tails in non-voiced segments captured by the recording device. From the figures, the discriminative characteristics of the environment, recording and playback devices are clearly visible in the replayed utterances. Speech utterances are generally characterized by 40:60 speech/non-speech ratio [23]. Since non-speech regions form a sizable part of an utterance, non-speech portions can be used to detect reverberation, and thus enable better detection of genuine and replayed utterances.

In this work, silence (S), voiced (V) and unvoiced (Uv) segments in a speech utterance are identified by using a voice activity detector with simple threshold on average energy, zero crossings [24] and spectral flatness [25]. The thresholds for average energy, zero-crossings and spectral flatness are chosen empirically using the development dataset. Replay detection is performed using the concatenated non-speech segments (S+Uv) of the original utterances.

IV. Decision Level Feature Switching

As shown in the Figures 1 and 2, the information in different replayed utterances varies based on the E-R-P configurations used to collect the replay samples. The proposed DLFS system tries to choose the best feature that better discriminates the genuine trial and the spoofed trial. Three features chosen for this purpose are: (i) MFCC (ii) CQCC and (iii) MFS. MFCC is a magnitude-based, commonly used feature that represents the gross characteristics of the vocal tract information [26]. CQCC proposed in [18] is based on Constant-Q Transform (CQT). This feature was shown to be robust against speech synthesis and voice conversion based spoofing attacks [27].

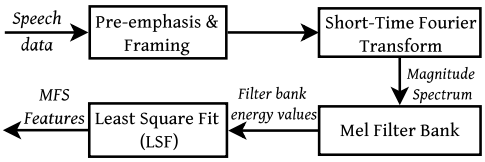


Fig. 3: Extraction of MFS features

MFS feature was proposed in [19] and extensively used in [28] for speaker verification. In this paper, MFS is used for the first time for detecting the replayed utterances. All these three features together will be referred to as "candidate features" in the following sections. Figure 3 shows the procedure for MFS feature extraction [29]. Slope calculated on the Mel-filterbank log energies is the measure of energy variations as a function of time which emphasize the formants.

As mentioned in Section I, three GMM systems are used as the baseline systems. In a baseline system, based on the log-likelihood ratio score ($\mathcal{S}_i^f$) between the genuine model and spoofed model of the f^{th} feature stream, every 'i'-th trial is tagged with a decision label ($\mathcal{L}_i^f$) , as in Equation 1.

$$\mathcal{L}_i^f = \begin{Bmatrix} \text{genuine}, & \text{if } \mathcal{S}_i^f > 0 \\ \text{spoof}, & \text{if } \mathcal{S}_i^f < 0 \end{Bmatrix} \quad (1)$$

where, $f \in \{\text{MFCC, CQCC, MFS}\}$. The DLFS uses $\{\mathcal{S}_i^f, \mathcal{L}_i^f\}$ pair of the feature-based baseline systems to identify the best feature of i-th trial. For every trial i, voting is performed on the labels $\mathcal{L}_i^f$ to get the final label as shown in Figure 4 and scores are compared according to Equation 1. Since the scores to be compared are from different feature spaces, the scores of the baseline systems are normalized using min-max normalization [30] as shown in Equation 2. The log-likelihood scores across the baseline systems may not be in same range. This means that the min_v and max_v values will be different for different baseline systems. The min-max score normalization transforms the baseline scores between the range $[min_v, max_v]$ to a common target range between $[min_t, max_t]$. In this paper, $[min_t, max_t]$ is empirically chosen as $[-1, +1]$ using development dataset.

$$\mathcal{S}_i^{f'} = (max_t - min_t) \times \frac{\mathcal{S}_i^f - min_v}{max_v - min_v} + min_t \quad (2)$$

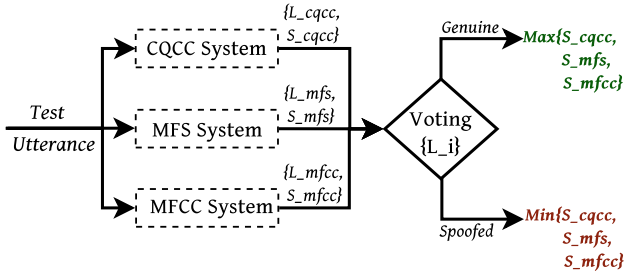


Fig. 4: Decision making in the proposed DLFS system. The dotted box in the figure corresponds to the testing process of different baseline systems.

If the voted label is *genuine*, the scores of the candidate feature $\mathcal{S}_i^f$ with the genuine label are compared, and the feature 'f' with *maximum* score is treated as the best feature for that trial. The score $\mathcal{S}_i^f$, corresponding to best feature 'f', is considered as the final score for that trial. If the voted label is *spoof*, then the scores of candidate features with the same label are compared. In this case, the *minimum*

TABLE I: EERs and Accuracy of the developed systems. For all these systems the training data alone is used to train the GMMs.

System Type	System Name	Features Used	Development Data				Evaluation Data			
			Full Utterance		non-voiced Utterance		Full Utterance		non-voiced Utterance	
			EER	Accuracy	EER	Accuracy	EER	Accuracy	EER	Accuracy
Baseline	CQC_BL	CQCC	**4.59**	82.57	**4.76**	89.94	34.86	20.80	37.49	23.18
	MFS_BL	MFS	6.89	**92.11**	5.98	**92.11**	**19.23**	**59.97**	**17.58**	**65.49**
	MFC_BL	MFCC	8.62	85.55	7.42	90.47	19.66	58.93	18.39	64.56
Score Fusion Systems	SF_CS	CQCC*MFS	6.40	90.90	4.48	91.63	28.61	31.73	22.50	58.92
	SF_CM	CQCC*MFCC	**4.62**	**92.69**	**3.39**	**92.74**	28.83	22.58	24.18	52.92
	SF_MS	MFCC*MFS	6.90	92.16	6.27	92.51	**19.32**	**61.31**	**17.50**	**67.11**
	SF_CMS	CQCC*MFCC*MFS	4.81	92.63	4.62	92.92	24.04	36.38	21.93	56.78
DLFS	DLFS_CS	CQCC+MFS	3.81	**96.55**	**2.99**	96.72	28.39	26.22	24.75	61.58
	DLFS_CM	CQCC+MFCC	**3.30**	96.54	3.12	**96.96**	27.71	28.83	21.80	63.58
	DLFS_MS	MFCC+MFS	6.68	93.15	6.62	93.74	**19.16**	**63.20**	**16.39**	**69.68**
	DLFS_CMS	CQCC+MFCC+MFS	4.04	92.69	3.92	95.73	21.28	54.16	19.36	67.57

score and the corresponding feature space is considered as the final score and best feature for the trial respectively. The maximum genuine score and minimum spoofed score implicitly represents the maximum discrimination between the genuine model and spoofed model, thereby increasing the confidence in classifying the trial.

V. Experimental Setup

All three candidate features are extracted with standard 25 ms window size and 10 ms window shift. Since channel information is crucial to discriminate a replayed utterance, all these features are extracted without any channel compensation techniques [31]. The feature specifications and GMM size are chosen empirically using the development and training dataset. Two sets of baseline systems are built in each feature space:

- System with whole utterance
- System with non-voiced utterance (Section III).

DLFS systems are implemented on the normalized scores of the baseline systems as explained in Section IV. Min-max normalization is performed on the scores to enable a fair comparison across the systems. Score fusion systems are also developed to compare the performance of the proposed systems. Score fusion paradigm uses information from more than one feature space by fusing the scores of the individual feature-based systems with a linear combination of weights. In this paper, the weights for the score fusion systems are learned by logistic regression from the training data, using Focal toolkit [33]. The details of the developed systems and corresponding performance are listed in Table I. When an *even* number of features are used as the candidate features, as in DLFS_CS, there are possibilities for the votes to be distributed equally for both genuine and spoofed classes. In such cases, the labels are chosen by comparing the absolute value of the maximum genuine score and minimum spoofed score. The class label with maximum magnitude, the corresponding feature, and score are considered as the final label, best feature, and final score for that trial respectively.

VI. Result Analysis

Equal Error Rate (EER) is chosen as the metric by the ASV-spoof team to evaluate the performance of the system [2]. The EERs and accuracy of all the developed systems are shown in Table I. The P_{miss} and P_{fa} are obtained using the *roch* function and EER is calculated using *roch2eer* function from the Bosaris toolkit [34]. The accuracies of the systems reported in the Table I are calculated as the ratio of the total number of hits to the total number of utterances in the dataset. CQCC-based baseline system (CQC_BL) performs better than MFCC- and MFS-based baseline systems on the development data in terms of EER. On the evaluation data, for the trials with unseen E-R-P conditions, MFS_BL system outperforms the other baseline systems.

TABLE II: Relative improvements (RI) of EER across systems.

Data Type	Best Baseline System		Best DLFS System		RI (in %)
	Full Utterance		Full Utterance		
	System ID	EER(%)	System ID	EER(%)	
Dev Data	CQC_BL	4.59	DLFS_CM	3.30	28.10
Eval Data	MFS_BL	19.23	DLFS_MS	19.16	0.36
	non-voiced Utterance		non-voiced Utterance		
Dev Data	CQC_BL	4.76	DLFS_CS	2.99	**37.18**
Eval Data	MFS_BL	17.58	DLFS_MS	16.39	7.26
	Best DLFS System (full utterance)		Best DLFS System (non-voiced utterance)		
Dev Data	DLFS_CM	3.30	DLFS_CS	2.99	**9.39**
Eval Data	DLFS_MS	19.16	DLFS_MS	16.39	**14.45**

In replayed utterances, the reverberation information and channel noise, induced due to the playback and replay devices usually have low-frequency components (2(b)), whereas the frequency components of the environment (ambient noise) may vary from high to low. CQCC gives high resolution for low-frequency components, and low resolution for high-frequency components [18] and hence, may not capture high-frequency ambient noise in the replayed utterances. This claim is evident from the accuracy of the replayed utterances in CQC_BL system (Table I). Even though the best proposed system does not surpass the state-of-the-art system in terms of the performance, the proposed DLFS system is less computationally intensive. The details of relative improvement of the proposed approach with the best baseline system and other score fusion systems in terms of EER are listed in Table II. Except for the CQC_BL system, the system with non-voiced utterance gives better performance than the system with the entire utterance, both in terms of accuracy and EER. Unlike the score fusion system, the proposed DLFS system do not require any weight learning process to learn the fusion weights. It is also evident from the results that the DLFS system outperforms all score fusion systems.

FFT-based MFCC and MFS features give uniform resolution to all frequency components, thereby capturing low-frequency as well as high-frequency channel noise to a certain extent. Formants are better represented by MFS than MFCC [19], [29]. Hence, MFS detects the replayed instances better, by capturing the reverberated remnant formant information in low energy frames. This is evident from the accuracies listed in Table I. Because of this varying representation of the various feature spaces, evaluating an utterance in the best feature space will be an appropriate countermeasure, to identify replayed utterances. Inferences from the results assure that the systems with MFS feature perform better than the rest of the systems. The systems developed from the non-voiced segments give better detection of replay utterances than the system with the entire utterance. Also, it is apparent from the accuracy that removal of voiced segments from the trials reduces the possibility of misclassification of the genuine and replayed utterances due to the spectral similarities between the voiced segments.

VII. Conclusion & Future work

A novel approach to replay attack detection using non-voiced segments is proposed in this paper. The remnant vocal tract information in the non voiced regions (owing to reflections from the environment) is exploited to detect whether a given utterance is replayed. This approach is evaluated on the ASV-spoof-2017 database using the best feature space of the utterances. The proposed DLFS system with non-speech utterance outperforms the baseline system with a relative improvement of 37.18% and 7.26% on the development and evaluation data respectively. The DLFS systems with non-voiced segments outperform the DLFS systems with the entire utterance on both development and evaluation data with a relative improvement of 9.39% and 14.45% respectively. From the results and inferences, we conclude that (i) the systems with MFS, outperform other systems in unseen E-R-P conditions and (ii) non-voiced regions alone are enough to distinguish a replayed instance from the corresponding genuine utterance.

References

[1] Zhizheng Wu et al., "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[2] Tomi Kinnunen et al., "Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof)," Feb 2017.

[3] Z. Wu et al., "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Dec 2014, pp. 1–5.

[4] Marcin Witkowski et al., "Audio Replay Attack Detection Using High-Frequency Features," in *INTERSPEECH*, Aug 2017, pp. 27–31.

[5] K. N. R. K. Raju Alluri et al., "SFF Anti-Spoofer: IIIT-H Submission for Automatic Speaker Verification Spoofing and Countermeasures Challenge 2017," in *INTERSPEECH*, Aug 2017, pp. 107–111.

[6] Parav Nagarsheth et al., "Replay attack detection using DNN for channel discrimination," *INTERSPEECH*, pp. 97–101, Aug 2017.

[7] Roberto Font et al., "Experimental analysis of features for replay attack detection–Results on the ASVspoof 2017 Challenge," *INTERSPEECH*, pp. 7–11, Aug 2017.

[8] Hemant A Patil et al., "Novel Variable Length Teager Energy Separation Based Instantaneous Frequency Features for Replay Detection," *INTERSPEECH*, pp. 12–16, Aug 2017.

[9] Galina Lavrentyeva et al., "Audio replay attack detection with deep learning frameworks," in *INTERSPEECH*, Aug 2017, pp. 82–86.

[10] Zhizheng Wu et al., "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, 2014, pp. 1–5.

[11] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar 2010, pp. 1678–1681.

[12] Jesús Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *Carnahan Conference on Security Technology*, Oct 2011, pp. 1–8.

[13] Jesús Villalba et al., "Detecting replay attacks from far-field recordings on speaker verification systems," in *Proceedings of the COST 2101 European Conference on Biometrics and ID Management*, Berlin, Heidelberg, 2011, pp. 274–285, Springer-Verlag.

[14] LP Zhang et al., "Prevention of impostors entering speaker recognition systems," *Journal of Tsinghua university (Science and Technology)*, vol. 48, no. S1, pp. 699–703, 2008.

[15] Javier Ramırez et al., "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3, pp. 271–287, 2004.

[16] T. Asha et al., "Feature switching in the i-vector framework for speaker verification," in *Interspeech*, Sep 2014, pp. 1125–1129.

[17] Saranya M. S. et al., "Feature-switching: Dynamic feature selection for an i-vector based speaker verification system," *Speech Communication*, vol. 93, pp. 53–62, 2017.

[18] Massimiliano Todisco et al., "A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients," in *The Speaker and Language Recognition Workshop, ODYSSEY*, June 2016.

[19] H. A. Murthy et al., "Robust text-independent speaker identification over telephone channels," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 554–568, Sep 1999.

[20] Kong-Aik Lee et al., "The reddots data collection for speaker recognition.," in *Interspeech*. 2015, pp. 2996–3000, ISCA.

[21] Zhizheng Wu and Haizhou Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5311–5327, 2016.

[22] N. Shokouhi et al., "Teager Kaiser Energy Operators for Overlapped Speech Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1035–1047, May 2017.

[23] F. Beritelli et al., "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 85–88, March 2002.

[24] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 478–482, Jul 2000.

[25] Yanna Ma et al., "Efficient voice activity detection algorithm using long-term spectral flatness measure," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 87, Jul 2013.

[26] K. S. R. Murty et al., "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, Jan 2006.

[27] Zhizheng Wu and others, "Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof)," Feb 2015.

[28] Srikanth Madikeri and H. A. Murthy, "Mel Filter Bank energy-based Slope feature and its application to speaker recognition," in *NCC*, Jan 2011, pp. 1–4.

[29] S. R. Madikeri et al., "Mel filter bank energy-based slope feature and its application to speaker recognition," in *NCC*, Jan 2011, pp. 1–4.

[30] Kevin L. Priddy et al., *Artificial Neural Networks: An Introduction Front Cover*, The International Society for Optical Engineers, Washington, 2005.

[31] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639–643, Oct 1994.

[32] Jiří Mekyska et al., *Score Fusion in Text-Dependent Speaker Recognition Systems*, pp. 120–132, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[33] Niko Brummer, "Focal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition scores," *Tutorial and User Manual. Spescom DataVoice*, 2007.

[34] Niko Brümmer et al., "The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF," 2013.