# A study on replay attack and anti-spoofing for text-dependent speaker verification

Zhizheng Wu*, Sheng Gao†, Eng Siong Chng‡ and Haizhou Li†,
*Centre for Speech Technology Research, University of Edinburgh, United Kingdom
†Human Language Technology Department, Institute for Infocomm Research, Singapore
‡School of Computer Enginnering, Nanyang Technological University, Singapore
Email: zhizheng.wu@ed.ac.uk

*Abstract*—**Replay, which is to playback a pre-recorded speech sample, presents a genuine risk to automatic speaker verification technology. In this study, we evaluate the vulnerability of text-dependent speaker verification systems under the replay attack using a standard benchmarking database, and also propose an anti-spoofing technique to safeguard the speaker verification systems. The key idea of the spoofing detection technique is to decide whether the presented sample is matched to any previous stored speech samples based a similarity score. The experiments conducted on the RSR2015 database showed that the equal error rate (EER) and false acceptance rate (FAR) increased from both 2.92 % to 25.56 % and 78.36 % respectively as a result of the replay attack. It confirmed the vulnerability of speaker verification to replay attacks. On the other hand, our proposed spoofing countermeasure was able to reduce the FARs from 78.36 % and 73.14 % to 0.06 % and 0.0 % for male and female systems, respectively, in the face of replay spoofing. The experiments confirmed the effectiveness of the proposed anti-spoofing technique.**

**Index Terms**: Speaker verification, spoofing attack, replay attack, anti-spoofing, countermeasure

## I. Introduction

Speaker verification is to automatically accept or reject an identity claim based on the provided speech sample. Typically, there are two types of systems: *text-dependent* and *text-independent*. Text-dependent speaker verification systems request the client to speak a promoted or fixed passphrase, while text-independent systems do not have such a constraint. As text-dependent speaker verification achieves high verification accuracy with short utterances, it is usually deployed for access control applications such as logic access control in the smartphone [1]. However, during deployment of the text-dependent systems, a major concern arises as whether such systems are still reliable in the face of spoofing attacks.

In the literature, there are four kinds of spoofing approaches [2], [3]: impersonation, speech synthesis, voice conversion and replay. Impersonation is an approach that an attacker tries to mimic a target genuine speaker and the vulnerability of speaker verification under impersonation attacks had been studied in [4], [5], [6], [7]. Speech synthesis is to generate the target genuine speaker's voice through a speech synthesis system to spoof the speaker verification systems [8], [9]. Voice conversion is to automatically manipulate an attacker's voice to mimic the target genuine speaker through a conversion function and the ability to spoof speaker verification systems had been studies in [10], [11], [12], [13], [14], [15], [16]. The last but the most easily implemented approach is replay, which was examined in [17], [18], [19], [20], [21] to assess the vulnerability of speaker verification. In this work, we focus on the replay attack and countermeasures.

Replay is a low technology spoofing attack approach without the need of speech processing techniques. Unfortunately there have been very few reported studies. The vulnerability of speaker verification to replay attack was evaluated in [17] for the first time. Pre-recorded isolated digits were concatenated and replayed to attack a hidden Markov model (HMM) based text-dependent speaker verification system. A considerable increase in both equal error rates (EERs) and false acceptance rates (FARs) was observed as a result of the replay attack. However, only two speakers' data were used in the database. The vulnerability of a text-independent joint factor analysis (JFA) system was evaluated in [18] and [19]. The pre-recorded speech samples were recorded through a far-field microphone and then replayed using a mobile phone. The experimental results also showed significant increase in the FAR as a result of replay attack. Note that only five speaker were involved in the dataset. Similar observation was also observed in [20], where a text-independent GMM-UBM system was adopted and a dataset collected from 13 speakers was employed.

Even though the previous work used several different speaker verification system, they concluded similar findings, that is significant increase in FARs as a result of replay attacks. Hence, the development of anti-spoofing technique to replay attacks is necessary to safeguard the speaker verification systems.

Let us start by reviewing the prior work on replay countermeasure studies. In [22], a replay attack detector was developed in the context of text-dependent speaker verification for the first time. The detector was designed to compare the verification sample with previous enrolled samples for enrolment or stored samples of past access attempts. The detector was evaluated in various playback detection tasks and was shown to achieve good performance in lowering the EERs. In [19], [23], a countermeasure was implemented to prevent replay attacks using far-field recordings. The countermeasure was designed based on the fact that the noise and reverberation levels will increase in the far-field recorded signals. In [20], an anti-spoofing approach based on examining the channel noised was proposed to protect a GMM-UBM speaker verification system, as the replayed speech contains the channel noise by two recording device and one loudspeaker used for replay, while licit recordings only have the channel noise introduced by the recording device of the speaker verification system.

The previous works have used much smaller datasets than those used in speaker verification. A typical speaker verification task usually involves several hundreds or even several thousands speakers. In this work, we first evaluate the vulnerability of the state-of-the-art text-dependent speaker verification in the face of replay attack. A standard database with a large number of speakers is adopted in the experiment. We then propose an anti-spoofing technique to detect the replay attacks to secure the speaker verification system. In the anti-spoofing technique, we extract some key points from the speech signal to compare the verification sample with the stored speech samples from enrolment or previous access attempts. If the similarity score is higher than a pre-defined threshold, we classify it as a replay attack sample; otherwise, classify it as a licit speech sample.

## II. Vulnerability of speaker verification to replay spoofing

### A. Speaker verification systems

In this work, we focus on the vulnerability of the text-dependent speaker verification systems. We employ a hierarchical acoustic modeling approach, which is presented in Figure 1. This structure consists of three layers: universal background model (top), text-dependent Gaussian mixture model (middle), and text-dependent hidden Markov model (bottom). The speaker- and text-dependent GMM is adapted from the UBM model using the maximum a posterior (MAP) adaptation [24], [25]. The speaker- and text-dependent HMM consists of five states, each of which is a GMM adapted from the GMM of the same text and speaker with the MAP criteria. Hence, the UBM and the GMM make the GMM-UBM system, and the UBM and the HMM make the HMM-UBM system. It is noted that the difference between GMM-UBM and HMM-UBM systems is whether temporal constraint is considered.

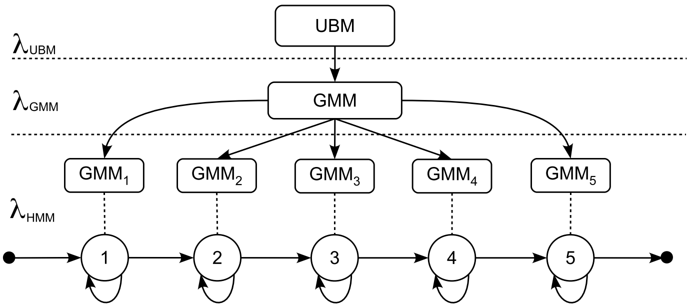

Fig. 1. An illustration of the hierarchical acoustic model approach to text-dependent speaker verification.

The GMM-UBM and HMM-UBM speaker verification systems make the verification decisions according to a log-likelihood ratio (LLR) score $\ell$, which is calculated as

$$\ell = \log \frac{p(\mathbf{X}|H_0)}{p(\mathbf{X}|H_1)}, \quad (1)$$

where $\mathbf{X}$ is the observation in the form of a collection of feature vectors, $H_0$ is the hypothesized speaker model, and $H_1$ is the alternative speaker model. In this work, the UBM model corresponds to the alternative speaker model, while speaker- and text-dependent GMM and HMM correspond to the hypothesized speaker model.

### B. Vulnerability of speaker verification

In speaker verification, three levels of features can be extracted to represent the speaker identity: a) short-term spectral and voice source features, such as Mel-frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs) and fundamental frequency (F0); b) spectro-temporal and prosodic features, such as modulation features and intonation patterns; c) high-level linguistic features, such lexical features [26]. These features can be used as the observations $\mathbf{X}$ for speaker verification.

In the context of replay attacks, the attacker plays the pre-recorded speech from the exact target speaker to spoof the speaker verification system. Hence, it is possible for the replay speech to have exactly the same spectral attributes, prosodic and high-level features as that of the target speaker. Figure 2 presents a comparison of a genuine speech and its corresponding replay speech. It is observed that the spectrogram of the replay speech is almost indistinguishable to the target genuine speech. If features are extracted from the replay spectrogram, it is possible to move the verification score towards that of the target speaker. In this way, the speaker verification system will lose the ability to distinguish genuine and impostor via replay.

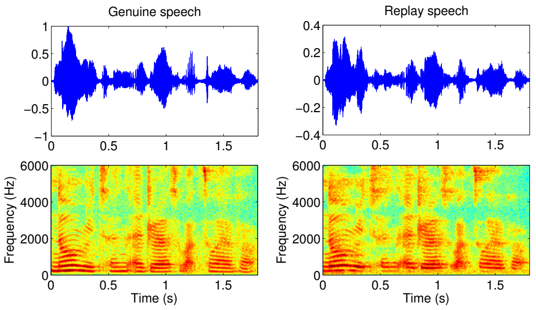

Fig. 2. Comparison of a genuine speech and its corresponding replay speech.

## III. Detection of replay attack

In response to the replay attack, a detection technique is proposed in this work, assuming that each passphrase spoken by the client to the speaker verification system has been recorded by the system. In this way, the detection technique is able to decide whether the verification sample is matched to any stored speech samples based a detection score. The detection score is calculated from two bitmaps, which relate to spectral peaks. We introduce the proposed detection technique in this section.

### A. Extraction of spectral bitmap

The proposed detection technique is based on the spectrogram bitmaps, which is similar to the audio fingerprint as proposed in [27]. Such fingerprint is found to be efficient in computation and robust to channel or noise in audio retrieval applications [27].

The spectral bitmap is related to spectral peaks, which are the time-frequency points having higher amplitudes than the neighboring points or than a pre-defined threshold. In this work, the spectral bitmap is computed as follows: a fast Fourier transform (FFT) is first applied to the speech signal to extract the magnitude spectrogram. Then, the spectrogram is divided into a number of non-overlapping blocks, each of which is a time-frequency block spanning a fixed range of frequencies and durations in frequency and time domains, respectively. After that, a mean and variance normalization is applied to the amplitude values in each block. Then, the normalized amplitude values are compared with a pre-defined threshold. If the amplitude of a time-frequency point is higher than the threshold, the point is chosen as a spectral peak and assigned a value of 1; otherwise, the point is not chosen as a spectral peak and assigned a value of 0. Finally, the new spectrogram consisting only values of 0 and 1 is generated as the spectral bitmap.

Figure 3 presents an example of spectral bitmaps of an original speech signal and its corresponding replay speech. From the observation of the spectral bitmap, it has some relations with the harmonics in the spectrogram, and the data point in the spectral bitmap is very sparse. In this way, the storage of the spectral bitmap does not require too much space.

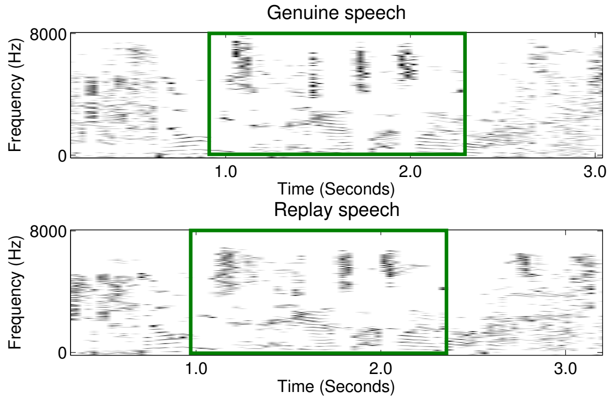

Fig. 3. Comparison of the bitmaps extracted from the genuine and replay speech signals.

### B. Calculation of detection score

The spectral bitmap derived as illustrated in Figure 3 is used to compute the replay detection score. In order to make it consistent with the speaker verification task, we define the replay speech which is a fake sample as an imposture sample, and the real speech sample as a genuine sample. In this way, the replay or imposture sample is expected to have a lower detection score than the real human speech sample.

To compute the detection score, the bitmap of the verification sample is compared with every stored bitmaps belonging to the target genuine speaker at runtime. An element-wise product is calculated between two bitmaps, and the summation of the element-wise product is used as the similarity score, which means the number of matched peak points between two bitmaps. A higher similarity score, much closer between two bitmaps. We define the detection score as the inverse of the similarity score for distinguishing the replay speech and the real speech.

As presented in Figure 3, a detection score can be calculated from the two blocks in the green color windows. If the two blocks match well, the detection score is extremely small, implying presence of replay speech; otherwise, the detection score is extremely high.

## IV. Experiments

### A. Designing the dataset

In this study, we used the Part I of the RSR2015 corpus [28], which is a standard benchmark database for text-dependent speaker verification, to design the dataset for replay attack and anti-spoofing. The database is recorded through multiple mobile devices. During the recording, each speaker reads 30 passphrases for each session, respectively, and each passphrase is repeated for nine sessions. More details of the corpus can be found in [28].

We divided the speakers into two non-overlapping sets: a background set consisting of 60 male and 60 female speaker, and an evaluation set including 30 male and 30 female speaker. The data from the speakers in the background set were used to train the UBM, and those in the evaluation set were used to adapt the UBM model as well as to evaluate the performance of the speaker verification systems.

As in the experiment, each utterance from each speaker in the evaluation set was used as a genuine trial for the same target model as well as an impostor trial against the other speakers of the same gender, this resulted into a huge number of trials. For the analysis purpose, we used only 10 utterances out of the 30 passphrases. For each utterance, three sessions, in particular sessions 1, 4 and 7, were used for enrolment to train the speaker- and text-dependent model, and the other six sessions were kept as verification trials. The six sessions were also replayed through a laptop and at the same time recorded by a laptop to produce the replay version of the natural human speech. In this work, we assumed the attackers know the gender of the target speaker as well as the promoted passphrase, in this way, we only considered the genuine and impostor trials that with matched passphase and gender. We note that the replay version of the target speaker's verification trial is used as the verification trial to match the exact target speaker's model, assuming the attacker has recorded the target speaker's previous verification samples.

Table I presents the statistics of the genuine, impostor and replay trials. In the experiments, we mixed the genuine trials and the impostor trials as a baseline test, and mixed the genuine trials and the replay trials as a replay test. Note that the genuine trials are exactly the same in the two tests, in this way, we are able to compare the error rates.

TABLE I
A SUMMARY OF GENUINE, IMPOSTOR AND REPLAY TRIALS USING RSR2015 DATABASE.

| | Male | Female | Total |
|---|---|---|---|
| Target speakers | 30 | 30 | 60 |
| Genuine trials | 1,796 | 1,797 | 3,593 |
| Impostor trials | 51,621 | 51,853 | 103,474 |
| Impostor trials via replay | 51,621 | 51,853 | 103,474 |

### *B. Experimental setups*

For feature representation in speaker verification, 12-order MFCCs with the delta and delta-delta coefficients were extracted from the speech signal at 16 kHz via a 27-channel Mel-frequency filter-bank. RASTA filtering, voice activation detection (VAD) to remove non-speech segments and sentence-level cepstral mean-variance normalization (CMVN) were performed on the 36 dimensional MFCCs.

### *C. Vulnerability evaluation of speaker verification*

In the first set of experiments, we evaluated the vulnerability of the speaker verification systems to replay spoofing. The equal error rate (EER) and false acceptance rate (FAR) results before and after the replay attack are presented in Table II. As a result of replay attacks, the EERs of the HMM-UBM system increase from 2.92 % and 2.39 % to 25.56 % and 20.05 % for male and female, respectively, and the EERs of the GMM-UBM system also increase considerably from 4.01 % and 3.67 % to 24.95 % and 21.95 % for male and female, respectively. Generally, from the EER results, the performance of the two systems are degraded considerably.

The FAR result is more related to spoofing attacks [16]. We calculated the FARs by setting the decision threshold at the EER point in order to compare the performance before and after spoofing. It is observed that after the replay attacks, the FARs of the HMM-UBM system increase to 78.36 % and 73.14 % for male and female, respectively, and the FARs of the GMM-UBM system also increase considerably, that is from 4.01 % and 3.67 % to 74.32 % and 65.28 % for male and female, respectively. Even though the performance of the HMM-UBM system is better than that of the GMM-UBM system in terms of EERs and FARs, under replay attacks, the two systems are both damaged and achieve similar performance.

The EERs and FARs reflect the underlying classifier scores shift before and after spoofing. We further took a look at the score distributions, which are presented in Figure 4. It is obviously observed that after replay spoofing, the impostor scores are moved towards the target genuine scores, and such shifting makes a considerable overlap between the impostor's score distribution and that of the genuine. This explains why the classifier is compromised in face of replay attacks.

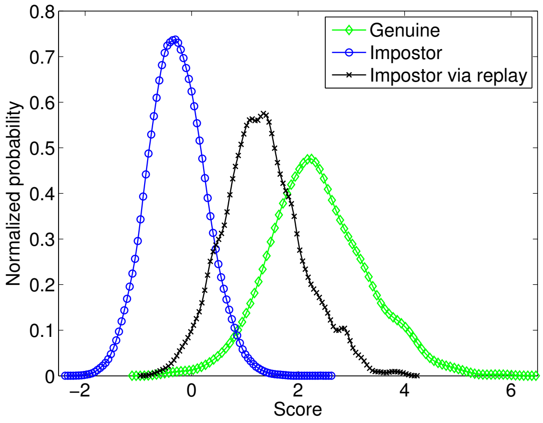

Fig. 4. Score distribution of the HMM-UBM system before and after replay attacks.

### *D. Spoofing countermeasure for speaker verification*

In the second set of experiments, we assessed the performance of the spoofing countermeasure for speaker verification. We first evaluated the performance of the stand-alone anti-spoofing replay detector. We made use of the trials for speaker verification evaluation to assess the performance of the replay detector. The original genuine and impostor trials were used as the natural human speech, while the replay trials as the replay speech. We also used the equal error rate to assess the performance, as there are two types of errors: replay speech is classified as human speech, and human speech is classified as replay. The countermeasure gave EERs of 0.00 % and 0.06 % for male and female, respectively. This confirms the effectiveness of the stand-alone detector.

We then evaluated the performance of the speaker verification systems with an integrated countermeasure. We set the decision thresholds for the speaker verification systems and the anti-spoofing detector to their own EER points. The false rejection rates (FRRs) and false acceptance rates (FARs) are presented in Table III. Comparing the performance of baseline and baseline+CM in Table II and III, respectively, it clearly shows that the performance of speaker verification systems is not affected by the countermeasure in face of zero-effort impostors. It is also observed that the countermeasure is able to reduce the FARs in face of replay attacks, by comparing the performance of replay and replay+CM in Table II and III, respectively.

## V. Conclusions

In this study, we evaluated the vulnerability of text-dependent speaker verification systems in the face of replay attacks, and also proposed a countermeasure for anti-spoofing in order to secure the speaker verification systems against replay attacks. The experiments confirmed the weakness of the speaker verification systems under replay spoofing, and

TABLE II
PERFORMANCE OF THE TEXT-DEPENDENT SPEAKER VERIFICATION SYSTEMS UNDER THE REPLAY SPOOFING ATTACK. THE FALSE ACCEPTANCE RATES (FARS) ARE OBTAINED BY SETTING THE THRESHOLD TO THE EQUAL ERROR RATE POINT ON BASELINE DATASET.

| Experiments | EER (%) | | | | FAR (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | HMM-UBM | | GMM-UBM | | HMM-UBM | | GMM-UBM | |
| | Male | Female | Male | Female | Male | Female | Male | Female |
| Baseline | 2.92 | 2.39 | 4.01 | 3.67 | 2.92 | 2.39 | 4.01 | 3.67 |
| Replay | 25.56 | 20.05 | 24.94 | 21.95 | 78.36 | 73.14 | 74.32 | 65.28 |

TABLE III
PERFORMANCE OF THE TEXT-DEPENDENT SPEAKER VERIFICATION SYSTEMS UNDER REPLAY SPOOFING ATTACK. THE FALSE REJECTION RATES (FRRS) AND FALSE ACCEPTANCE RATES (FARS) ARE OBTAINED BY SETTING THE THRESHOLD TO THE EQUAL ERROR RATE (EER) POINTS OF THE BASELINE DATASET. CM = COUNTERMEASURE

| Experiments | FRR (%) | | | | FAR (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | HMM-UBM | | GMM-UBM | | HMM-UBM | | GMM-UBM | |
| | Male | Female | Male | Female | Male | Female | Male | Female |
| Baseline + CM | 2.90 | 2.39 | 4.01 | 3.67 | 2.92 | 2.39 | 4.01 | 3.67 |
| Replay + CM | 2.90 | 2.39 | 4.01 | 3.67 | 0.06 | 0.00 | 0.06 | 0.00 |

also confirmed the effectiveness of the proposed anti-spoofing detector, which decides whether the verification sample is matched to any previous stored speech sample through a similarity/detection score.

## REFERENCES

[1] K. A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," in *IEEE Signal Processing Society Speech and language Technical Committee Newsletter,*, 2013.

[2] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech*, 2013.

[3] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. DeLeon, "Voice anti-spoofing," in *Handbook of biometric anti-spoofing*, S. Marcel, S. Z. Li, and M. Nixon, Eds. Springer, 2014.

[4] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. Int. Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004.

[5] Y. Lau, D. Tran, and M. Wagner, "Testing voice mimicry with the YOHO speaker verification corpus," in *Knowledge-Based Intelligent Information and Engineering Systems*. Springer, 2005, pp. 907–907.

[6] J. Mariéthoz and S. Bengio, "Can a professional imitator fool a GMM-based speaker verification system?" IDIAP Research Report (No. Idiap-RR-61-2005), 2005.

[7] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry," in *Proc. Interspeech*, 2013.

[8] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 2001.

[9] 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," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 8, pp. 2280–2290, 2012.

[10] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Interspeech*, 2007.

[11] 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 *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[12] 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 *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2012.

[13] F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[14] Z. Wu, A. Larcher, K. A. Lee, E. S. Chng, T. Kinnunen, and H. Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints," in *Proc. Interspeech*, 2013.

[15] Z. Kons and H. Aronowitz, "Voice transformation-based spoofing of text-dependent speaker verification systems," in *Proc. Interspeech*, 2013.

[16] Z. Wu and H. Li, "Voice conversion and spoofing attack on speaker verification systems," in *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2013.

[17] J. Lindberg, M. Blomberg *et al.*, "Vulnerability in speaker verification-a study of technical impostor techniques," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 1999.

[18] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA 10 workshop*, 2010, pp. 131–134.

[19] ——, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Biometrics and ID Management*, ser. Lecture Notes in Computer Science, C. Vielhauer, J. Dittmann, A. Drygajlo, N. Juul, and M. Fairhurst, Eds. Springer, 2011, pp. 274–285.

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

[21] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Proc. Int. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2014.

[22] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[23] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *IEEE Int. Carnahan Conf. on Security Technology (ICCST)*, 2011.

[24] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[25] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, January 2000.

[26] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[27] A. Wang, "An industrial strength audio search algorithm," in *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, 2003, pp. 7–13.

[28] A. Larcher, K.-A. Lee, B. Ma, and H. Li, "RSR2015: Database for text-dependent speaker verification using multiple pass-phrases." in *Proc. Interspeech*, 2012.