Multimed Tools Appl (2016) 75:5311–5327
DOI 10.1007/s11042-015-3080-9

On the study of replay and voice conversion attacks to text-dependent speaker verification

Zhizheng Wu[1] · **Haizhou Li**[2]

Received: 3 August 2015 / Revised: 3 October 2015 / Accepted: 13 November 2015 /
Published online: 3 December 2015

Abstract Automatic speaker verification (ASV) is to automatically accept or reject a claimed identity based on a speech sample. Recently, individual studies have confirmed the vulnerability of state-of-the-art text-independent ASV systems under replay, speech synthesis and voice conversion attacks on various databases. However, the behaviours of text-dependent ASV systems have not been systematically assessed in the face of various spoofing attacks. In this work, we first conduct a systematic analysis of text-dependent ASV systems to replay and voice conversion attacks using the same protocol and database, in particular the RSR2015 database which represents mobile device quality speech. We then analyse the interplay of voice conversion and speaker verification by linking the voice conversion objective evaluation measures with the speaker verification error rates to take a look at the vulnerabilities from the perspective of voice conversion.

Keywords Speaker verification · Spoofing attack · Replay · Voice conversion · Security

1 Introduction

Automatic speaker verification (ASV) [17, 28] is a low-cost biometric solution. Unlike other forms of biometrics, such as fingerprint or iris recognition, a speech sample can be acquired remotely using existing landline, cellular and voice-over IP communication channels

✉ Zhizheng Wu
zhizheng.wu@ed.ac.uk

Haizhou Li
hli@i2r.a-star.edu.sg

[1] The Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, UK

[2] Human Language Technology Department, Institute for Infocomm Research (I2R), Singapore, Singapore

without additional hardware. ASV technology has already been deployed in applications like real-time caller verification [6, 34] and smartphone login [27] to validate transactions in e-commerce and to safeguard personal information.

ASV systems typically operate in one of the two input modes: *text-independent* or *text-dependent*. Text-independent mode assumes free text, while text-dependent one enforces or prompts the user to speak a given pass-phrase. Text-independent systems are often used for off-line screening, indexing or forensic uses, involving non co-operative users. They can be used to verify customer's identity from free-worded conversion for example in call centres. Text-dependent systems requiring co-operative users [5, 13], in turn, are commonly used for authentication applications in view of their relative high recognition accuracy.

It is understood that there is no absolutely secure biometrics, and a biometric authentication system can always be intentionally circumvented or spoofed [36]. Since ASV usually takes place without a face-to-face contact with a human operator, spoofing to ASV becomes a fundamental concern when deploying a system. As reviewed in [45], there are at least four types of spoofing attacks: impersonation, replay, speech synthesis and voice conversion. Among them, replay, speech synthesis and voice conversion are three more effective spoofing attacks. However, past work generally focuses on a specific spoofing attacks, and makes the comparison to be difficult. In our previous work [47], speech synthesis and voice conversion attacks have been analysed and compared using the same database in the context of text-independent ASV. In this work, we focus on replay and voice conversion attacks in the context of text-dependent ASV.

1.1 Related work

Impersonation by human beings as a natural way to spoof ASV systems has received attention in [10, 12, 24]. This attack has occasionally been successful in spoofing speaker verification systems. However, impersonators tend to mimic prosody, pronunciation and lexicon rather than the spectral cues used by ASV systems. Apparently, there are other more consistent ways of attack as facilitated by recent advances in speech processing to spoof ASV systems.

Speech synthesis which generates speech with a decent target speaker voice quality presents an emerging threat to the security of speaker verification systems. Having enough technical skills, one can easily produce speaker adapted voices using tools such as Festival[1]. Indeed, unit selection [14], statistical parametric [54] and hybrid [35] synthesis methods are able to generate speech adapted to a target speaker with an acceptable quality. Generally speaking, a modern statistical parametric synthesis technique first trains an average voice model from large corpus, which is subsequently adapted to a specific target speaker using a small amount of adaptation utterances [51, 53]. Although speech synthesis has been shown to increase the error rates of state-of-the-art systems to unacceptable levels in [8, 31, 32, 38, 47], it is not straightforward to perform spoofing, as it requires text input.

Aside from impersonation and speech synthesis, *replay* – the rendering of previously recorded target speaker utterances [30, 43] – might be the most common spoofing technique to ASV, as it does not require the attackers to have any speech technology knowledge. Although such attack might not be effective in generating utterances for specific content to maintain a live conversation in call-centre applications, it is still one of the most effective attacks against authentication systems which use fixed pass-phrase.

[1] http://festvox.org/index.html

Voice conversion [41] that offers another effective way to generate synthetic speech with a decent voice quality, attempts to achieve the same effect as human impersonation and adapted speech synthesis, but operates on a speech signal itself. Most voice conversion techniques do not require transcriptions, prosody prediction, or additional off-line corpora. During the execution of spoofing, voice conversion is hand-free without requiring any additional efforts by human. The past individual studies [2–4, 15, 18, 20, 33, 48] have shown that voice conversion techniques are able to increase the error rates of state-of-the-art classifiers to unacceptable levels.

1.2 Motivation and contributions

A recent review [45] highlights that a) lacking of standard databases makes the comparison across spoofing types difficult; and b) development of protocols and countermeasures for speaker verification lags behind that for other biometric systems, even though there is increasingly accumulating work towards developing countermeasures [1, 44, 52] and their integrations with speaker verification system [16, 48]. In this work, we will be a step closer to better understanding how spoofing attacks and speaker verification performance are interrelated, which can be useful for designing spoofing protocols or evaluation metrics. Our contributions are three-fold.

Firstly, we attempt to analyse the spoofing effects of voice conversion and replay attacks using the same protocol, and evaluate vulnerability of text-dependent systems on the RSR2015 corpus [22, 23]. This is the first step to make a standard text-dependent spoofing database that includes multiple spoofing types. With this protocol, we provide a detailed look into the vulnerability of ASV systems, and compare the effectiveness of different spoofing attacks to text-dependent speaker verification.

Secondly, we study the interplay of voice conversion quality and speaker recognition performance. We note that *speaker similarity* is central to both applications; measuring it is the sole task of speaker verification systems but it also finds use in objective evaluations of the performance of voice conversion methods. Even if speaker verification systems are occasionally used as 'black-box' evaluators of speaker similarity in voice conversion studies, they are massively data-driven complex systems and require keeping some data for universal background modelling (UBM) [37] or other uses. Consequently, speaker similarity in voice conversion systems is usually assessed through direct acoustic distortion measurement between the converted and the target features, enabling a convenient and inexpensive procedure to optimise parameters of a new voice conversion technique or to compare different voice conversion methods. But from the perspective of spoofing attacks, the relevant question is whether acoustic distortion is a useful predictor of the false acceptance rate under spoofing attacks?

We follow the standard speaker verification architecture which is supposed to take only natural voice as input, and add a voice conversion system or a replay mechanics at the input point to create spoofing attacks. In a genuine trial, a genuine voice goes directly to the feature extraction module, while in an impostor trial, an impostor's voice passes through the voice conversion to impersonate the target genuine speaker.

2 Vulnerability of speaker verification to attacks

When deploying a speaker verification system, the system is expected to be accurate to regular clients, and also robust against spoofing attacks. As pointed out in [11, 45], spoofing

attacks can take place at two locations in a speaker verification system: at the microphone sensor and during the transmission of the acquired speech signal. At the sensor level, an impostor, also called an adversary, could compromise the system by replaying a pre-recorded speech signal or impersonating the target speaker at the sensor. During the transmission, the acquired speech signal could be replaced by a falsifying one. Generally, a spoofing attack is to employ a forged signal as the system input in either of the above two locations.

A typical speaker verification system is optimised to accept genuine speakers and to reject impostors, assuming natural human speech. Speech consists of three primary constituents: voice timbre, prosody and language content. Speakers can therefore be characterised at three different levels [17, 29]: a) short-term spectrum; b) prosody; and c) high-level idiolectical/lexical features. Being information-rich and practical to compute, spectral features — usually, the Mel-Frequency Cepstral Coefficients (MFCCs) [7] — are the primary features used by modern recognisers, and prosodic features may be added to further enhance accuracy [9, 19, 39].

As discussed in [50], feature extraction in speaker verification is one of the weak links. In a replay attack, an attacker plays a pre-recorded speech from the exact target speaker to spoof a text-dependent speaker verification system. Hence, it is possible for the replayed speech to have exactly the same spectral attributes, prosody and high-level lexical features as that of the target speaker, presuming the text-dependent speaker verification system uses fixed pass-phase. Figure 1 presents a comparison of a genuine speech and its corresponding replayed speech. It is observed that it is hard to distinguish the spectrograms between the genuine and replayed speech. if spectral features are extracted from the replayed spectrogram, it is possible to achieve a high verification score that matches the target speaker, and hence the speaker verification system will lose the ability to prevent replay attacks.

Voice conversion operates on voice timbre and prosody in order to mimic a target speaker's voice. It is hence able to move an impostor's spectral feature and prosodic feature distributions towards those of a genuine target speaker's distribution. Utilising pairs of

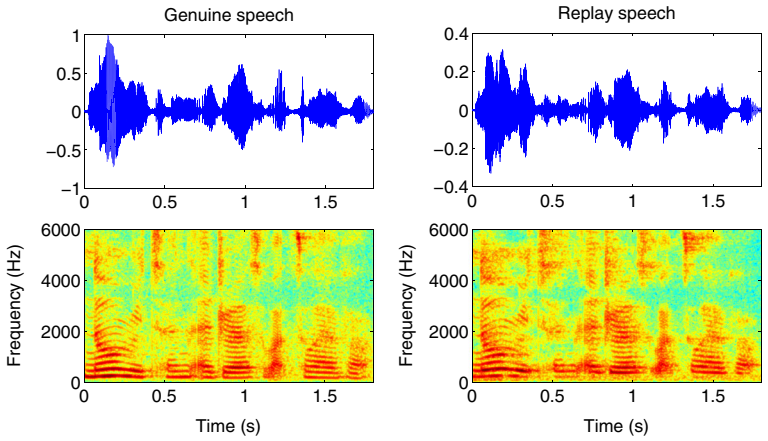

Fig. 1 Comparison of a genuine speech and its corresponding replay speech. It is hard to distinguish between genuine and replay speech from the time-domain and spectrum-domain representations. The Figure is adopted from [46]

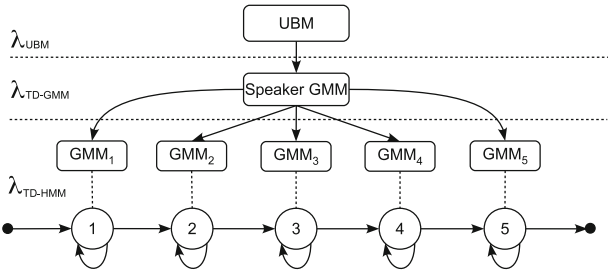

Fig. 2 Hierarchical acoustic modelling approach for text-dependent speaker verification

training vectors $\{(x_i, y_i)\}$ from the source and the target speakers, a linear or nonlinear mapping function $y = F(x)$ is then trained to approximate the distribution of target speaker for any new inputs x, presenting a great risk to speaker verification systems that utilise similar features. We hence argue that short-term spectral and prosodic features are two weak links of a verification system facing voice conversion attacks.

In this work, we use joint density Gaussian mixture model (JD-GMM) and a harmonic plus noise model (HNM) based vocoder[2] to perform voice conversion. Mel-Cepstral coefficients (MCCs) are converted using JD-GMM described in [49] while F0 is converted by equalizing the means and variances of source and target speakers in log-scale.

3 Speaker verification systems

In this study, we investigate the effect of spoofing attacks against text-dependent systems. To allow a tractable analysis, we use a hierarchical acoustic modeling [21, 26] as shown in Fig. 2, in which two variants of text-dependent speaker models are progressively trained from the same universal background model, according to the formulation of the maximum a posterior (MAP) adaptation [25, 37].

We consider two text-dependent classifiers:

- **TD-GMM**: In this setup, a speaker- and text-dependent GMM model is adapted from a universal background model (UBM). The top and middle layers in Fig. 2 hence correspond to the good old GMM-UBM [37]. In practice, GMMs are pass-phrase dependent, so we refer to this GMM-UBM model as *GMM-based text-dependent model* (TD-GMM).
- **TD-HMM**: In the bottom layer, a speaker-dependent and sentence-level hidden Markov model (HMM) is adapted from the middle layer TD-GMM. In particular, each state of the HMM is a GMM adapted from the TD-GMM of the speaker by using the MAP criteria. We hence call this pass-phrase and speaker-dependent HMM approach as *HMM-based text-dependent model* (TD-HMM).

We consider the structure in Fig. 2 to present TD-GMM and TD-HMM for two reasons. Firstly, the RSR2015 text-dependent speaker verification database consists of utterances with very short duration (3 seconds of nominal speech, see Section 4.1). For short-duration training and test utterances, the conventional GMM-UBM with MAP adaptation has shown

[2] http://aholab.ehu.es/ahocoder/

to perform equally well as compared to JFA or PLDA [40]. Second, a more tractable analysis is possible given that the bottom layer HMM models additional temporal information absent in the middle layer GMM.

The likelihood ratio of TD-GMM is calculated between λ_{UBM} and $\lambda_{\mathrm{TD-GMM}}$, and, similarly, the likelihood score of TD-HMM is obtained between λ_{UBM} and $\lambda_{\mathrm{TD-HMM}}$.

All the speaker verification systems use the same acoustic front-end consisting of 12 MFCCs with delta and delta-delta coefficients computed via 27-channel mel-frequency filterbank. RASTA filtering, voice activity detection (VAD) and utterance cepstrum mean-variance normalization (CMVN) are employed as postprocessing. The VAD decisions of test segments are derived from the original baseline datasets.

4 Database and protocol

4.1 Spoofing datasets

In light of the mass market adoption of speaker verification technology in smartphone [27], we decide to focus on mobile device quality speech. We use the nine sessions of the first two parts of the RSR2015 database [22]. This corpus has been recorded using multiple mobile devices and smartphones over nine recording sessions and this corpus can be used as a standard benchmark database for text-dependent speaker verification system development and evaluation. During the recording, a speaker reads 30 pass-phrases for each session of Part I and 30 short commands for each session of Part II. The average duration of the pass-phrases is 3.2 seconds. Two non-overlapping sets of speakers are defined: a background set including 60 male and 60 female speakers, and an evaluation set of 30 male and 30 female speakers. Speakers from the background set are reserved for training a universal background model (UBM) [37] needed for constructing our classifiers.

For the experiments, each speaker from the evaluation set is used both as a target speaker and as an impostor against other speakers of same gender. Out of the 9 sessions available for each speaker, three sessions are used for enrolment (sessions 1, 4 and 7) while the six remaining sessions are used as test materials. Note that enrolment and test sessions are defined so that the recording device used for verification test is different from the one used during the enrolment. To avoid overlapping between speaker model training and conversion function training, we further split the 30 sentences into two groups. Pass-phrases 1 to 10 are used for speaker verification experiments while sentences 11 to 30 are set aside for training the voice conversion function. Thus, 60 utterances from each speaker are used to produce genuine and impostor trials (10 pass-phrases and 6 sessions). The statistics of the trials are presented in Table 1. Given this protocol, we note that only the genuine and impostor trials

Table 1 Statistics of the baseline and spoofing datasets from RSR2015 database (VC=voice conversion)

	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
Impostor trials via VC	51,621	51,853	103,474

with matched pass-phrase and matched gender are considered. That is, the attacker knows the prompted pass-phrase.

To produce replay trials, the six genuine sessions used as testing materials were replayed through a laptop and at the same time recorded by a laptop to produce the replayed version of the genuine speech. In this work, we assume the attackers know the gender information of the target speaker and can obtain the prompted pass-phrase. In this way, we only considered the genuine and impostor trials that with matched pass-phase 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. We also note that for different attackers to spoof the same target speaker, we repeated the same replay speech, and this explains why the number of replay trials is the same as that of impostor trials.

To generate the voice conversion spoofing datasets, we pass the test samples for the impostor trials through voice conversion while keeping the genuine trials untouched. This allows us to focus solely on the effects of spoofing attack. We design the spoofing attack datasets by repeating the following three steps for each impostor trial:

– Estimate a conversion function between an impostor and a target genuine speaker's speech;
– Employ the conversion function to modify each test sample of the impostor;
– Adopt the converted speech sample as a testing sample of the impostor.

In practice, we use the JD-GMM method to generate the voice conversion spoofing dataset, and make the number of converted trials be the same as that of the original impostor trials.

We pool the genuine trials and original impostor trials as a baseline test, and at the same time pool the genuine trials and impostor trials via replay or voice conversion as a spoofing attack test. We expect to see the decisions of genuine trials remain the same between the baseline and the spoofing test, and an increase of false alarm arising from the converted speech samples. In this way, we are able to compare the performance and examine the spoofing attack effect. The actual numbers of trials are presented in Table 1.

4.2 Performance evaluation measures

A speaker verification trial where the test and enrolment utterances share the same speaker identity variable is a *genuine* trial; otherwise, we call it an *impostor* trial. Given a test sample, the acceptance or rejection decision made by a verification system falls into one of the four groups shown in Table 2, where false acceptance and false rejection are the misclassifications. Often classifier parameters are optimized to obtain low *equal error rate* (EER), corresponding to a verification threshold at with false acceptance rate (FAR) and false rejection rate (FRR) are equal.

Table 2 Four groups of trial decisions in speaker verification

	Decision	
	Acceptance	Rejection
Genuine test	Correct acceptance	False rejection
Impostor test	False acceptance	Correct rejection

In a spoofing attack scenario, a speaker verification system is unaware of the attack and is deployed with a fixed threshold, assuming that the testing samples are natural human voices. FAR is therefore a natural criterion for evaluating vulnerabilities of a speaker verification system under spoofing attack. Formally, let $\text{FAR}(\theta, \mathcal{D})$ and $\text{FRR}(\theta, \mathcal{D})$ denote FAR and FRR, evaluated at operating point (threshold) θ on dataset (corpus) $\mathcal{D}$. Let $\mathcal{D}_{\text{base}}$ denote a baseline corpus consisting of genuine and zero-effort impostor trials (i.e. no dedicated spoofing attempts). Further, let $\mathcal{D}_{\text{attack}}$ be a corpus that shares the same genuine trials as $\mathcal{D}_{\text{base}}$ but in which *all* the impostor trials have been replaced by voice conversion samples simulating a dedicated attack. With these notations, our protocol is:

1. Determine EER threshold on $\mathcal{D}_{\text{base}}$:
$$\theta_{\text{EER}} = \arg\min_{\theta} |\text{FRR}(\theta, \mathcal{D}_{\text{base}}) - \text{FAR}(\theta, \mathcal{D}_{\text{base}})|$$
2. Compute $\text{EER}(\theta_{\text{EER}}, \mathcal{D}_{\text{attack}})$ and $\text{FAR}(\theta_{\text{EER}}, \mathcal{D}_{\text{attack}})$ and observe their relative increase w.r.t. baseline dataset.

5 Experimental results and analysis

The objective of our experiments is to evaluate the vulnerabilities of automatic speaker verification systems under replay and voice conversion spoofing. As the study involves both speaker verification, replay and voice conversion techniques, we look into the research problem from two different angles: a) examining the performance of speaker verification systems under spoofing attacks; and b) analysing the effectiveness of replay and voice conversion as a spoofing approach. In this work, details of the speaker verification systems, such as feature extraction and speaker modelling techniques, are assumed unknown to an attacker.

In this section, we present three case studies. The first case study evaluates the vulnerability of two variants of text-dependent ASV systems under the same replay attack, while in the second case study, we examine the vulnerabilities of the same two variants of text-dependent systems under voice conversion attacks, and also study how the number of JD-GMM training utterances affects the outcomes. In the third case study, we extend the second case study by having an inquiry into the relationship between voice conversion performance and the respective effects of spoofing attacks to gain further insights.

5.1 Case study 1: Overall effect of replay attack to speaker verification accuracy

In the first set of experiment, we examined the vulnerability of ASV systems to replay attack. EER and FAR results before and after replay spoofing are shown in Table 3. In the face of replay, the EERs of the HMM-UBM systems increases from 2.92 % and 2.39 % to 25.56 % and 20.05 % for male and female speakers, respectively, and that of the GMM-UBM systems also increase considerably from 4.01 % and 3.67 % to 24.94 % and 21.95 % from male and female, respectively. In general, the performance in terms of EERs of the two systems is degraded considerably. This observation is consistent with previous studies over relatively smaller datasets.

FARs are more related to spoofing attacks, and hence we calculated FARs by setting the decision threshold at the EER point in order to compare the performance before and after spoofing. The FARs are presented in Table 3. It is observed that in the face of replay spoofing, 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.

Table 3 Performance of *text-dependent* speaker verification systems under voice conversion and replay spoofing attacks

Spoofing	EER (%)				FAR (%)			
	TD-HMM		TD-GMM		TD-HMM		TD-GMM	
	Male	Female	Male	Female	Male	Female	Male	Female
None (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
VC-2	3.90	1.78	5.90	3.98	4.80	1.06	9.12	4.30
VC-5	5.07	2.51	8.24	5.84	9.17	2.64	16.94	8.43
VC-10	7.04	2.82	11.28	6.88	16.20	3.77	26.60	11.19
VC-20	8.30	3.12	13.34	7.31	21.87	4.68	33.23	13.20

False acceptance rate (FAR) is obtained by setting the threshold to the equal error rate point on baseline dataset. Assuming the impostor knows the exact pass-phrases

Even though the performance of the HMM-UBM system is better than that of the GMM-UBM system in terms of EERs and FARs, the two systems are both damaged and achieve similar performance under the same replay spoofing.

We further take a look at the classifier score distributions before and after replay spoofing, as the increase of EERs and FARs reflects the shift of underlying classifier scores as a result of replay spoofing. The score distributions of the HMM-UBM system before and after replay spoofing are presented in Fig. 3. It is clearly observed that as a result of the replay spoofing, the impostor score distribution is moved towards that of the target genuine scores, and such score shifting makes a considerable overlap between the impostor's score distribution and that of the target genuine. This phenomenon also explains the reason why the ASV systems are compromised in the face of replay attacks.

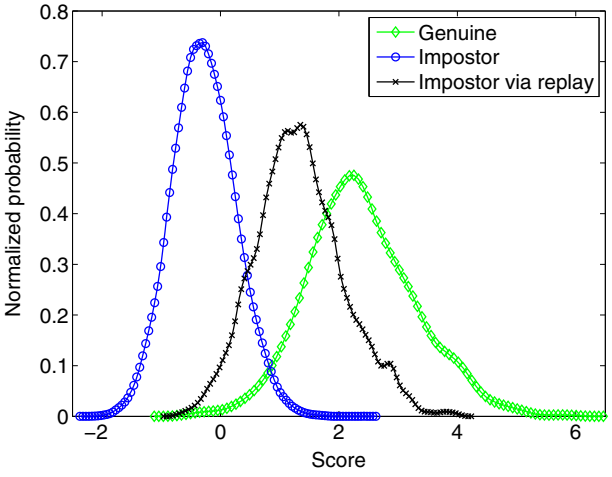

Fig. 3 Score distributions of the male HMM-UBM system before and after replay attacks

5.2 Case study 2: Overall effect of voice conversion attacks to speaker verification accuracy

In the second case study, we examine the vulnerabilities of two variants of text-dependent systems, and also study how the number of JD-GMM training utterances affects the outcomes. In the experiments, we vary the number of parallel training utterances from 2 to 20. In particular, we used 2, 5, 10 and 20 utterances, respectively, to estimate the conversion function. They are labelled as VC-2, VC-5, VC-10 and VC-20 in Table 3, and the corresponding number of Gaussian components in JD-GMM is empirically set to 4, 8, 16 and 32, respectively.

We first compare the EERs of TD-HMM speaker verification systems before and after voice conversion attack. As shown in Table 3, before spoofing attack, the EERs of TD-HMM systems are 2.92 % and 2.39 % for male and female, respectively. As a result of spoofing attack using 2 utterances for training, the EER of male speakers increases to 3.90 %, however, the EER of female speakers *decreases* to 1.78 %. This might be because of too few voice conversion training utterances. Indeed, when we increase the number of parallel training utterances, the EERs increase to 8.30 % and 3.12 % for male and female, respectively. When more than 5 utterances are used to estimate the conversion function, the EERs after spoofing attack are higher than that before spoofing attack for both male and female.

We then evaluate the performance of TD-GMM speaker verification systems. Before spoofing attack, the EERs are 4.01 % and 3.67 % for male and female, respectively, which are slightly higher than that of TD-HMM systems. As presented in Table 3, even when only 2 utterances are used for estimating the voice conversion function, EERs increase over baseline for both genders. When using 20 utterances to estimate a conversion function, the EERs increase to 13.34 % and 7.31 % for male and female, respectively. We note that for both TD-HMM and TD-GMM systems, male speakers have higher EERs than those of female speakers.

The FAR results of both TD-HMM and TD-GMM are also presented in Table 3. When using 20 utterances to train a conversion function, the FARs of the TD-HMM verification systems increase from 2.93 % and 2.39 % of baseline to 21.87 % and 4.68 % after spoofing attack for male and female, respectively. The FARs of the TD-GMM systems increase from 4.01 % and 3.67 % before spoofing attack to 33.23 % and 13.20 % after spoofing attack for male and female, respectively. We observe a similar effect for EERs as the number of training utterances varies.

To sum up, increasing the number of training utterances for voice conversion increases both EERs and FARs. Voice conversion with enough training training data is hence able to move an impostor's feature distribution towards that of a target speaker, and presents an increased threat to both TD-HMM and TD-GMM verification systems. Since TD-HMM uses hidden Markov model to capture both feature distribution and temporal sequence information, it outperforms TD-GMM system which only models the feature distribution even under voice conversion attack. Note that the temporal sequence information remains the same as in the original impostor samples after voice conversion.

As a further analysis, we present the score distributions of the TD-HMM system for males and females in Figs. 4 and 5, respectively. We observe a similar score shifting pattern for TD-GMM. Trials on the right hand side of the decision threshold are falsely accepted while those on the left hand side are correctly decided as rejected. For male speakers, all the four cases of spoofing attacks consistently move the imposture scores towards the right side of the decision threshold. As for female speakers, we have a similar observation except the case of two utterances (VC-2) which slightly shifts the score distribution towards the

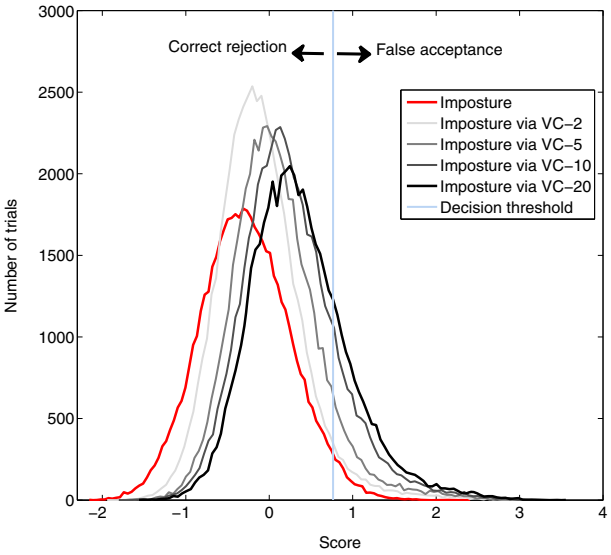

Fig. 4 Score distributions of the male TD-HMM system before and after spoofing

left with a reduced score variance. Using more VC training data (VC-5, VC-10 and VC-20), score distribution translates to the right, consistent with the FAR results in Table 3. We note that VC-2 shows unsuccessful attack when the threshold is set to EER point, but it might be effective for other settings of the decision threshold. For instance, a system optimized to have a lower false rejection rate (FRR).

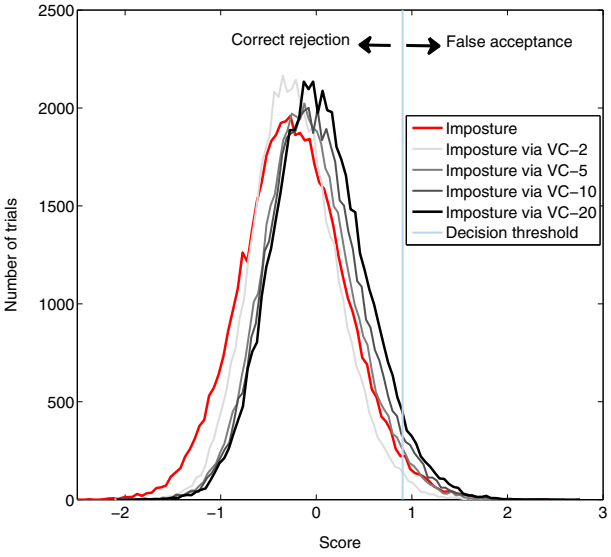

Fig. 5 Score distributions of the female TD-HMM system before and after spoofing

Generally, when enough training data is used to estimate the conversion function, voice conversion spoofing attack is able to compromise both TD-HMM and TD-GMM verification systems. This confirms the risk of voice conversion spoofing attack and the vulnerabilities of both TD-HMM and TD-GMM verification systems.

5.3 Case study 3: Voice conversion performance vs spoofing effect

In the third case study, we extend the vulnerability evaluation study by having an inquiry into the relationship between voice conversion performance and the respective effects of spoofing attacks against speaker verification to gain further insights. From the previous text-dependent analysis, we have the following observations: a) the female systems have lower EERs and FARs than the male systems, b) it shows unsuccessful spoofing attack simulated by voice conversion using only two utterances on the female TD-HMM system, c) the more training data for voice conversion, the higher EERs and FARs of both TD-HMM and TD-GMM under spoofing attacks.

In voice conversion, spectral distortion, in particular Mel-cepstral distortion (MCD) [42], is frequently used as an objective evaluation measure to predict voice conversion performance. MCD is calculated between the source or converted speech and reference target speech to measure the distance between two speech signals, as follows

$$\text{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(c_d - c'_d)^2}, \tag{1}$$

where D is the dimension of MCC feature, c_d is the d-th dimension reference target feature, and c'_d is the d-th dimension source feature with or without conversion. A lower MCD value indicates higher similarity of the compared speech signals. In voice conversion, the objective is to minimise the MCD between the source and target features. A perfect voice conversion system will be able to achieve a MCD result of zero.

From the perspective of EER and FAR, a lower MCD value may indicate higher EER and FAR, as higher similarity between two speech signals implies a more difficult classification task. To calculate spectral distortion, we randomly select 5,000 source-target utterance pairs for each gender. We note that RSR2015 pass-phrase part is a *parallel* dataset, that is, each speaker speaks the same utterances. We use dynamic time warping (DTW) to perform optimal frame alignment between source and target utterances to get the frame pairs for calculating MCD. To make a fair comparison, the converted utterance shares the alignment information with source utterances. Hence, the spectral distortion of a source utterance with and without conversion to a target utterance is comparable. The calculation is done frame-by-frame and we report the average distortion.

Figure 6a-c presents the comparison of between spectral distortions and FARs. Without voice conversion, the spectral distortions between source and target speech are 7.81 dB and 8.07 dB for male and female, respectively. A higher spectral distortion of female implies larger variability across speakers, therefore, it is easier to classify female speakers and is more difficult to estimate conversion functions for female speaker conversion. Refer to Table 3, the female TD-HMM and TD-GMM systems have lower EERs than that of the male systems. Due to the difficulty in capturing the large variability across female speaker, it hence is hard to build a "good" conversion function using limited training data. This explains why using only two utterances as training data does not increase the spoofed FAR for female case.

When we increase the number of training utterances, spectral distortions decrease from 6.86 dB and 7.09 dB of two utterances to 6.46 dB and 6.79 dB of 20 utterances for male

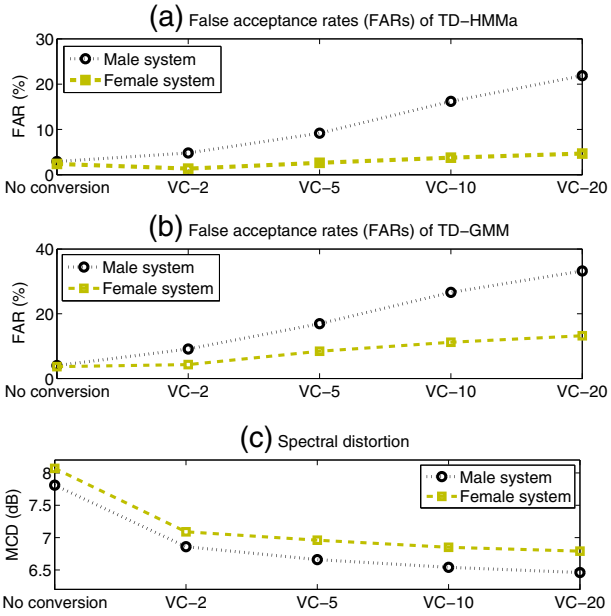

Fig. 6 Comparison across spectral distortions, FARs of TD-HMM systems and that of TD-GMM systemss on the RSR2015 dataset. **a**: FARs of male and female TD-HMM systems. No conversion means before spoofing attack. VC-2, VC-5, VC-10 and VC-20 represent different voice conversion attacks. **b**: FARs of male and female TD-GMM systems. **c**: Spectral distortions of male and female voice conversion systems

and female, respectively. Instead, FARs of both TD-HMM and TD-GMM increase as we expected. This changing trends between spectral distortions and FARs go opposite directions. We also note that for each conversion case, female speakers always have a higher spectral distortion than that of male speakers. This phenomenon is observed in the EER and FAR results, where female speakers always achieve lower EERs and FARs than those of male speakers, due to the reason we explain above.

6 Conclusions

We have examined the vulnerabilities of text-dependent speaker verification systems under replay and voice conversion attacks, and also established a link between voice conversion quality (Mel-Cepstral distortions) and spoofing success (false acceptance rates). The experimental results confirmed the vulnerabilities of text-dependent systems. Our main findings are:

1. Text-dependent ASV systems are vulnerable to both replay and voice conversion attacks. The first finding is expected and is consistent with previous studies on text-independent ASV.
2. HMM-based text-dependent systems in which temporal speech information matters were found more resistant in the face of voice conversion spoofing than systems lacking temporal modelling, while under replay attacks, the experimental results show that HMM-based systems with temporal modelling is equally vulnerable to GMM-based systems without temporal modelling.

3. Successful voice conversion attacks to HMM-based text-dependent classifiers require sufficiently many training utterances – in our findings, five or more for the set-up considered. To attack the female TD-HMM system, it was observed that the spoofing was unsuccessful when only two voice conversion training utterances were used (VC-2), however, the effect disappears with increased number of VC training utterances. With too few VC training utterances, the speaker transformation (JD-GMM) might be undertrained (e.g. not stable full covariance matrices to formulate the transformation function), causing conversion artefacts that the TD-HMM correctly considers being far off from natural speech, leading improved separation of genuine and (converted) impostor score distributions.
4. The vulnerabilities of speaker verification systems in terms of FARs have a high correlation with the voice conversion performance. This finding is naturally expected. Speaker verification is to find a decision boundary between impostor and target genuine speakers' feature distributions to make the decision. Voice conversion might be able to move the impostor's feature distribution to cross the decision boundary, as the objective of voice conversion is to shift a source speaker's feature distribution to match that of a target speaker, that is to minimise the MCD.

Our findings suggest that there are two possible directions to enhance the performance under spoofing: (1) implementing stand-alone countermeasures as a complementary component to speaker verification systems. The countermeasures could be motivated by the artefacts introduced in the voice conversion process. The second direction is to (2) improve the fundamentals of speaker verification, such as including time sequence information and high level features. As voice conversion is not perfect, the impostor's feature distribution is not exactly matched with that of the target speaker, it would be interesting to involve converted speech to improve the speaker modelling techniques.

References

1. Alegre F, Amehraye A, Evans N (2013) A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In: Proceedings of the international conference on biometrics: theory, applications and systems (BTAS)
2. Alegre F, Vipperla R, Evans N, et al. (2012) Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals. In: Proceedings interspeech
3. Bonastre JF, Matrouf D, Fredouille C (2006) Transfer function-based voice transformation for speaker recognition. In: Proceedings Odyssey: the speaker and language recognition workshop
4. Bonastre JF, Matrouf D, Fredouille C (2007) Artificial impostor voice transformation effects on false acceptance rates. In: Proceedings interspeech
5. Campbell J (1997) Speaker recognition: A tutorial. Proc IEEE 85(9):1437–1462
6. Center ST VoiceGrid (TM) RT: Sophisticated distributed solution for real-time speaker identification. In: http://speechpro.com/product/biometric/voicegridrt
7. Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 28 (4):357–366
8. De Leon P, Pucher M, Yamagishi J, Hernaez I, Saratxaga I (2012) Evaluation of speaker verification security and detection of HMM-based synthetic speech. IEEE Trans Audio Speech Lang Process 20(8):2280–2290
9. Dehak N, Dumouchel P, Kenny P (2007) Modeling prosodic features with joint factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 15(7):2095–2103
10. Farrús M, Wagner M, Anguita J, Hernando J (2008) How vulnerable are prosodic features to professional imitators? In: Proceedings Odyssey: the speaker and language recognition workshop
11. Faundez-Zanuy M, Hagmüller M, Kubin G (2006) Speaker verification security improvement by means of speech watermarking. Speech Comm 48(12):1608–1619

12. Hautamäki RG, Kinnunen T, Hautamäki V, Leino T, Laukkanen AM (2013) I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In: Proceedings interspeech
13. Hebert M (2008) Text-dependent speaker recognition. In: Benesty J, Sondhi M, Huang Y (eds) Springer Handbook of Speech Processing. Springer Berlin, Heidelberg, pp 743–762
14. Hunt AJ, Black AW (1996) Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP)
15. Jin Q, Toth A, Black A, Schultz T (2008) Is voice transformation a threat to speaker identification? In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP)
16. Khoury E, Kinnunen T, Sizov A, Wu Z, Marcel S (2014) Introducing i-vectors for joint anti-spoong and speaker verication. In: Proceedings interspeech
17. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: From features to supervectors. Speech Comm 52(1):12–40
18. Kinnunen T, Wu Z, Lee K, Sedlak F, Chng E, Li H (2012) Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP)
19. Kockmann M, Burget L, Cernocky J (2010) Investigations into prosodic syllable contour features for speaker recognition. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP)
20. Kons Z, Aronowitz H (2013) Voice transformation-based spoofing of text-dependent speaker verification systems. In: Proceedings interspeech
21. Larcher A, Bonastre JF, Mason JS (2013) Constrained temporal structure for text-dependent speaker verification. Digital Signal Processing 23(6):1910–1917
22. Larcher A, Lee KA, Ma B, Li H (2012) The RSR2015: Database for text-dependent speaker verification using multiple pass-phrases. In: Proceedings interspeech
23. Larcher A, Lee KA, Ma B, Li H (2014) Text-dependent speaker verification: Classifiers, databases and RSR2015. Speech Comm 60:56–77
24. Lau YW, Wagner M, Tran D (2004) Vulnerability of speaker verification to voice mimicking. In: Proceedings of the IEEE international symposium on intelligent multimedia, video and speech processing
25. Lee CH, Huo Q (2000) On adaptive decision rules and decision parameter adaptation for automatic speech recognition. Proc IEEE 88(8):1241–1269
26. Lee KA, Larcher A, Thai H, Ma B, Li H (2011) Joint application of speech and speaker recognition for automation and security in smart home. In: Proceedings interspeech
27. Lee KA, Ma B, Li H (2013) Speaker verification makes its debut in smartphone. In: IEEE signal processing society speech and language technical committee newsletter
28. Li H, Ma B (2010) Techware: Speaker and spoken language recognition resources [best of the web]. IEEE Signal Proc Mag 27(6):139–142
29. Li H, Ma B, Lee KA (2013) Spoken language recognition: From fundamentals to practice. Proc IEEE 101(5):1136–1159
30. Lindberg J, Blomberg M, et al. (1999) Vulnerability in speaker verification-a study of technical impostor techniques. In: Proceedings of the European conference on speech communication and technology (Eurospeech)
31. Masuko T, Hitotsumatsu T, Tokuda K, Kobayashi T (1999) On the security of HMM-based speaker verification systems against imposture using synthetic speech. In: Proceedings of the European conference on speech communication and technology (Eurospeech)
32. Masuko T, Tokuda K, Kobayashi T (2000) Imposture using synthetic speech against speaker verification based on spectrum and pitch. In: Proceedings of the international conference on spoken language processing (ICSLP)
33. Matrouf D, Bonastre JF, Fredouille C (2006) Effect of speech transformation on impostor acceptance. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing(ICASSP)
34. Nuance: Nuance voice biometrics. In: http://www.nuance.com/landing-pages/products/voicebiometrics/
35. Qian Y, Soong FK, Yan ZJ (2013) A unified trajectory tiling approach to high quality speech rendering. IEEE Trans Audio Speech Lang Process 21(2):280–290
36. Ratha NK, Connell JH (2001) Bolle, R.M.: Enhancing security and privacy in biometrics-based authentication systems. IBM Syst J 40(3):614–634
37. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted gaussian mixture models. Digital signal processing 10(1):19–41
38. Satoh T, Masuko T, Kobayashi T, Tokuda K (2001) A robust speaker verification system against imposture using a HMM-based speech synthesis system. In: Proceedings of the European conference on speech communication and technology (Eurospeech)

39. Shriberg E, Ferrer L, Kajarekar S, Venkataraman A, Stolcke A (2005) Modeling prosodic feature sequences for speaker recognition. Speech Comm 46(3):455–472
40. Stafylakis T, Kenny P, Ouellet P, Perez J, Kockmann M, Dumouchel P (2013) Text-dependent speaker recognition using PLDA with uncertainty propagation. In: Proceedings interspeech
41. Stylianou Y, Cappé O, Moulines E (1998) Continuous probabilistic transform for voice conversion. IEEE Transactions on Speech and Audio Processing 6(2):131–142
42. Toda T, Black AW, Tokuda K (2007) Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans Audio Speech Lang Process 15(8):2222–2235
43. Villalba J, Lleida E (2010) Speaker verification performance degradation against spoofing and tampering attacks. In: Proceedings FALA 10 workshop
44. Wu Z, Chng E, Li H (2012) Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In: Proceedings interspeech
45. Wu Z, Evans N, Kinnunen T, Yamagishi J, Alegre F, Li H (2015) Spoofing and countermeasures for speaker verification: a survey. Speech Comm 66:130–153
46. Wu Z, Gao S, Cling ES, Li H (2014) 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 ASC)
47. Wu Z, Khodabakhsh A, Demiroglu C, Yamagishi J, Saito D, Toda T, King S (2015) SAS: A speaker verification spoofing database containing diverse attacks. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP)
48. Wu Z, Kinnunen T, Chng E, Li H, Ambikairajah E (2012) A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case. In: Proceedings Asia-Pacific signal information processing association annual summit and conference (APSIPA ASC)
49. Wu Z, Larcher A, Lee KA, Chng ES, Kinnunen T, Li H (2013) Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints. In: Proceedings interspeech
50. Wu Z, Li H (2014) Voice conversion versus speaker verification: an overview. APSIPA Transactions on Signal and Information Processing 3(e17). doi:10.1017/ATSIP.2014.17
51. Wu Z, Swietojanski P, Veaux C, Renals S, King S (2015) A study of speaker adaptation for DNN-based speech synthesis. In: Proceedings interspeech
52. Wu Z, Xiao X, Chng ES, Li H (2013) Synthetic speech detection using temporal modulation feature. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP)
53. Yamagishi J, Kobayashi T, Nakano Y, Ogata K, Isogai J (2009) Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans Audio Speech Lang Process 17(1):66–83
54. Zen H, Tokuda K, Black AW (2009) Statistical parametric speech synthesis. Speech Comm 51(11):1039–1064

Zhizheng Wu is a research fellow in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh since 2014, and he received the Ph.D. degree from Nanyang Technological University (NTU), Singapore. He initiated and organised the first Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) as a special session at Interspeech 2015. He received the best paper award in APSIPA ASC 2012. His research interests includes speech synthesis, voice conversion, spoofing and anti-spoofing, and speaker verification.

Haizhou Li received the B.Sc, M.Sc, and Ph.D degrees in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990 respectively. He joined the Institute for Infocomm Research in Singapore in 2003, where he is now the Research Director of the Institute, the Principal Scientist and the Department Head of Human Language Technology. He is also an adjunct Professor of the Department of Electrical and Computer Engineering at the National University of Singapore, the School of Computer Engineering at Nanyang Technological University; and the School of Electrical Engineering and Telecommunications at the University of New South Wales, Australia.

Dr. Li has worked on speech and language technology in academia and industry since 1988. He has taught in The University of Hong Kong (1988-1989), South China University of Technology in Guangzhou, China (1990-1994), Nanyang Technological University in Singapore (since 2006), and University of Eastern Finland (2009). He was a Visiting Professor at CRIN/INRIA in France (1994-1995). Prior to joining I2R, he was a Research Manager in Apple-ISS Research Centre (1996-1998), Research Director of Lernout & Hauspie Asia Pacific (1999-2001), Vice President of InfoTalk Corp. Ltd (2001-2003). He co-founded Baidu-I2R Research Centre in Singapore (2012). Dr. Li was known for his technical contributions to several award-winning speech products, such as Apple's Chinese Dictation Kits for Macintosh (1996) and Lernout & Hauspie's Speech-Pen-Keyboard Text Entry Solution for Asian languages (1999). He was the architect of a series of major technology deployments that include TELEFIQS voice-automated call centre service in Singapore Changi International Airport (2001), voiceprint engine for Lenovo A586 Smartphone (2012), and Baidu Music Search (2013).

Dr. Li is currently the Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING (2015-2017). He has served as an Associate Editor (2008-2012) and Senior Area Editor (2014-2016) of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, Associate Editor (2012-2013) of ACM TRANSACTIONS ON SPEECH AND LANGUAGE PROCESSING , Computer Speech and Language (2012-2015), and Springer International Journal of Social Robotics (2008-2012), and a Member of IEEE Speech and Language Processing Technical Committee (2013-2015). He is the Vice President of the International Speech Communication Association (ISCA, 2013-2014), the President of Asia Pacific Signal and Information Processing Association (APSIPA, 2015-2016), the President of the Chinese and Oriental Language Information Processing Society (COLIPS, 2011-2013), an Executive Board Member of the Asian Federation of Natural Language Processing (AFNLP, 2006-). Dr. Li served as the Local Arrangement Chair of SIGIR 2008 and ACL-IJCNLP 2009. He was appointed the General Chair of ACL 2012 and INTERSPEECH 2014, and Technical Program Chair of ISCSLP 1998, APSIPA Annual Summit and Conference 2010, IEEE Spoken Language Technology Workshop 2014, and IEEE ChinaSIP 2015.

Dr. Li was the recipient of National Infocomm Awards 2002, Institution of Engineers Singapore (IES) Prestigious Engineering Achievement Award 2013, President's Technology Award 2013, and MTI Innovation Activist Gold Award 2015 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by Nokia Foundation and an IEEE Fellow in 2014 for leadership in multilingual, speaker and language recognition. Dr. Li is a member of ACL, ACM, IEICE and ISCA.