SAS : A SPEAKER VERIFICATION SPOOFING DATABASE CONTAINING DIVERSE ATTACKS

Zhizheng Wu[1] *Ali Khodabakhsh*[2] *Cenk Demiroglu*[2] *Junichi Yamagishi*[1,3]
Daisuke Saito[4] *Tomoki Toda*[5] *Simon King*[1]

[1]University of Edinburgh, United Kingdom [2]Ozyegin University, Turkey
[3]National Institute of Informatics, Japan [4]University of Tokyo, Japan
[5]Nara Institute of Science and Technology, Japan

ABSTRACT

This paper presents the first version of a speaker verification spoofing and anti-spoofing database, named **SAS** corpus. The corpus includes nine spoofing techniques, two of which are speech synthesis, and seven are voice conversion. We design two protocols, one for standard speaker verification evaluation, and the other for producing spoofing materials. Hence, they allow the speech synthesis community to produce spoofing materials incrementally without knowledge of speaker verification spoofing and anti-spoofing. To provide a set of preliminary results, we conducted speaker verification experiments using two state-of-the-art systems. Without any anti-spoofing techniques, the two systems are extremely vulnerable to the spoofing attacks implemented in our **SAS** corpus.

Index Terms— Database, speaker verification, spoofing attack, security, speech synthesis, voice conversion

1. INTRODUCTION

In the past decade or so, automatic speaker verification (ASV) technology has advanced significantly, to the point of mass market adoption [1]. A major concern when deploying an ASV system is whether the system is secure against spoofing attacks. Recently, an increasing number of studies have assessed the vulnerability of ASV systems to various forms of spoofing attacks [2, 3, 4], including impersonation [5, 6], replay [7], speech synthesis [8, 9] and voice conversion [10, 11, 12, 13, 14]. Efforts have also been made to develop individual countermeasures to protect ASV systems against specific spoofing attacks. However, the lack of a *standard spoofing database* is holding back the development of more general countermeasures [15, 2].

In the literature, we find that the design of a spoofing database depends very much on the particular spoofing approach assumed in each specific study, and this has resulted in a diverse set of individual spoofing databases, none of which is helpful for developing generalised countermeasures. In [5], an impersonation database was designed based on the YOHO speaker verification database [16], whereas the authors in [6] designed a small impersonation database in Finnish independently of any prior database. For speech synthesis-based spoofing [8, 9], the publicly-available Wall Street Journal (WSJ) database has been employed to generate spoofing materials. The clean recording conditions of WSJ enable advanced speech synthesis techniques to be applied. Conversely, standard NIST speaker recognition evaluation (SRE) databases have been used to construct voice conversion-based spoofing databases [10, 11, 12, 17]. The NIST databases include channel noise, which present substantial challenges to the more advanced forms of speech synthesis or voice conversion. In [14], a non-publicly available database was employed to design a voice conversion spoofing database for text-dependent speaker verification. As we can see, all of the databases we can find are focused on one specific variety of spoofing. This makes comparisons across different spoofing approaches difficult (e.g., is voice conversion better than state-of-the-art speech synthesis?), and generalised countermeasures (e.g., the detection of non-human speech) cannot readily be developed or evaluated using these databases.

There are a few attempts to design spoofing databases involving multiple varieties of spoofing attacks. In [7, 18], a spoofing database was designed based on RSR2015 [19] including both replay and voice conversion attacks. However, only a simple voice conversion technique was used. In [17], voice conversion, speech synthesis and artificial signal spoofing approaches were implemented on the NIST 2006 subset. However, only one voice conversion and one speech synthesis approach was employed, and only male speakers were included. No standard spoofing database exists that includes a diverse variety of spoofing techniques. Such a database is needed for conducting repeatable and comparable spoofing attack studies and to drive the development of generalised countermeasures that are effective across a wide variety of spoofing methods.

The ideal standard spoofing database should include all available spoofing approaches. As pointed out in [15, 3], a publicly available spoofing database and a competitive challenge based on such a common database are needed for spoofing and countermeasure (also known as anti-spoofing) research.

In this paper, we report our progress in developing such a standard spoofing and anti-spoofing database involving multiple varieties of spoofing attacks, for both text-dependent and text-independent scenarios. We present the current *spoofing and anti-spoofing* (SAS) database – that is the **SAS** corpus – and a preliminary set of benchmark results, for text-independent ASV. The database includes both speech synthesis and voice conversion spoofing attacks, which are two of the most accessible and effective spoofing approaches currently available [2, 3]. To improve the diversity of the data (and therefore the generalisation ability of countermeasures developed using it), we employed one speech synthesis techniques in two training scenarios and seven voice conversion techniques in one training scenario. We use state-of-the-art statistical parametric speech synthesis to implement speech synthesis spoofing, while the voice conversion spoofing sets were created using one publicly-available open-source toolkit and six state-of-the-art conversion techniques.

To the best of our knowledge, this is the first attempt to include such a diverse range of spoofing attacks in a single database. The **SAS** corpus will be *publicly available at no cost and we welcome* *additions to it from other researchers*.[1] In this paper, we present benchmark results when attacking two state-of-the-art speaker verification systems.

2. PROTOCOL

The **SAS** spoofing database starts with the Voice Cloning Toolkit (VCTK) database[2] from the Centre for Speech Technology Research (CSTR), which is English and freely available. The VCTK database was recorded in a hemi-anechoic chamber using an omni-directional head-mounted microphone (DPA 4035) at a sampling rate of 96 kHz. The motivation for starting with clean studio-recorded speech is that it allows for spoofing attacks that rely on such data. Channel and noise factors can always be simulated at a later date, but in this paper we focus only on spoofing under clean conditions.

To design the spoofing database, we took speech data from VCTK which comprises 45 male and 61 female speakers, and downsampled the signals to 16 kHz at 16 bits-per-sample. We also divided the data from each speaker into five parts:

- **Part-A:** 24 parallel utterances (i.e., same across all speakers) per speaker: training data for spoofing.
- **Part-B:** 20 non-parallel utterances per speaker: additional training for spoofing.
- **Part-C:** 50 non-parallel utterances per speaker: enrolment data for client model training in speaker verification.
- **Part-D:** 100 non-parallel per speaker: development set for speaker verification.
- **Part-E:** Around 200 non-parallel utterances per speaker: evaluation set for speaker verification.

We note that in Part-C, Part-D, and Part-E, all the sentences are randomly selected from newspapers without any repeating sentence across all speakers.

2.1. Speaker verification enrolment and evaluation

We first introduce the protocol for standard speaker verification evaluation. The enrolment data of each client was selected from Part-C under two scenarios: 5-utterance or 50-utterance enrolments. For 5 utterances this means around 5 to 6 seconds, and for 50 utterances around 1 minute of speech material.

The development set was created from Part-D. It involves genuine trials and impostor trials. All utterances from a client speaker in Part-D were used as genuine trials, and this results in 4500 male and 6100 female genuine trials. For the impostor trials, 10 randomly selected non-target speakers were used as impostors. All Part-D utterances from a specific impostor were used as impostor trials against the client's model, leading to 45000 male and 61000 female impostor trials. This set is aimed at tuning the system and deciding thresholds.

The evaluation is drawn from Part-E. In a similarly fashion to the development set, we generated 9446 male and 13385 female genuine trials, and 85592 male and 118000 female impostor trials. This set is for assessing the performance of speaker verification systems. A summary of the development and evaluation sets is shown in Table 1.

Table 1. Number of trials in the development and evaluation sets.

	Development			Evaluation		
	Male	Female	Total	Male	Female	Total
Target speakers	45	61	106	45	61	106
Genuine trials	4500	6100	10600	9446	13385	22831
Impostor trials	45000	61000	106000	85592	118000	203592
Spoofed trials	45000	61000	106000	85592	118000	203592

2.2. Spoofing preparation and execution

We now introduce the protocol for producing the spoofing materials. We designed two training sets: *small* and *large*. The small set consists of data only from Part-A, while the large set includes data from both Part-A and Part-B. Would-be attackers should select one of these to train their spoofing system. The small set comprises parallel training data, and so enables attackers to use voice conversion methods reliant on parallel training data, such as the method implemented in Festvox.[3]

In **SAS** , during the execution of speech synthesis spoofing, the transcript of an impostor trial was used as the textual input to the speech synthesis systems, while for voice conversion (VC) spoofing, the speech signal of the impostor trial was the input to the VC system. As a result, the zero-effort impostor trial, the speech synthesis spoofed trial and the voice conversion spoofed trial all have the same language content (i.e., word sequence).

The spoofing systems were used to generate spoofing materials for both development and evaluation, and so the number of spoofed trials is exactly the same as the number of impostor trials (Table 1). This allows fair comparisons to be made between non-spoofed and spoofed speaker verification results.

2.3. Evaluation metric

As discussed above, the protocol for speaker verification follows the NIST SRE style, so the evaluation metric designed for NIST evaluation can be easily adopted. For example, the performance measures Equal Error Rate (EER), False Acceptance Rate (FAR), False Rejection Rate (FRR) and Detection Cost Function (DCF) can be applied. In the benchmarking results we present here, EERs and FARs will be reported.

3. SPOOFING APPROACHES

In the current version of **SAS** , spoofing materials comprise the output from two speech synthesis systems and seven voice conversion systems. These systems are built using both open-source software and our internal systems. Next, we briefly describe the systems that were used to generate the spoofing materials in **SAS** .

NONE: This is a baseline zero-effort impostor trial in which the impostor's own speech is used directly with no attempt to match the target speaker.

SS-SMALL: This HMM-based TTS system is based on the statistical parametric speech synthesis framework described in [20]. The speaker adaptation techniques in this framework allow the generation of a synthetic voice using as little as a few minutes of recorded speech from the target speaker, making it an effective and easily-accessible tool for SV spoofing. We used the latest version (2.2) of the open-source code "HTS" [21, 22].

In the speech analysis and the average voice training phase, the STRAIGHT vocoder with mixed excitation is used, which results in

[1] **SAS** is being released under a Creative Commons licence. More details are available here: https://wiki.inf.ed.ac.uk/CSTR/SASCorpus

[2] http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

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

60-dimension Bark-Cepstral coefficients, log F_0 and 25-dimension band-limited aperiodicity measures [23, 24]. Hidden semi-Markov models (HSMMs) [25] are trained on a large multi-speaker database called voice bank corpus [26] that include hundreds of English speakers to simultaneously model acoustic features and duration. In the speaker adaptation phase, the speaker-independent HSMMs are transformed using structural variational Bayesian linear regression [27] followed by MAP, using the target speaker's data from Part-A. Both the output probability density functions for the acoustic features and the duration model parameters are transformed. To synthesise speech, acoustic feature parameters are generated from the adapted HSMMs using a parameter generation algorithm that considers global variance [28]. An excitation signal is generated using mixed excitation and pitch-synchronous overlap and add [29] and used to excite a Mel-logarithmic spectrum approximation (MLSA) filter [30] corresponding to the STRAIGHT Bark cepstrum, to create the final synthetic speech waveform.

SS-LARGE: This system is the same as SS-SMALL, except that a larger set of adaptation data comprising both Part-A and Part-B was used when adapting the speaker-independent HSMMs to each target speaker.

VC-FESTVOX: This is the voice conversion toolkit within the publicly-available open-source Festvox system. It is based on the algorithm proposed in [31], which is a joint density Gaussian mixture model with maximum likelihood parameter generation considering global variance. We used the Part-A (i.e., small) set of parallel training data, and kept the default settings of the toolkit, except that the number of Gaussian components in the mixture distributions was set to 32.

VC-GMM: This is another standard GMM-based voice conversion method also using the parallel training data from Part-A. It is very similar to VC-FESTVOX but with some enhancements. STRAIGHT was used as the speech analysis-synthesis method to extract high-quality speech parameters, such as F_0, spectral envelope, and aperiodicity measures. The search range for F_0 extraction was automatically optimized speaker by speaker to reduce errors. A power threshold for extracting active frames used to estimate the joint density GMM was also optimized automatically per speaker. Two GMMs were trained for separately converting the 1st through 24th Mel-Cepstral coefficients (MCCs) and 5 band aperiodicity measures. The number of mixture components was set to 32 for the spectral features and 8 for the aperiodicity measures, respectively. For some speaker pairs, the number of components was reduced when defunct mixture components were automatically removed. To enhance the variance of the converted spectral parameter trajectories, GV-based post-filtering [32] was used instead of GV-based parameter conversion.

VC-KPLS: This voice conversion system uses kernel partial least square (KPLS) regression [33], trained on the Part-A (small) parallel data. 300 reference vectors and a Gaussian kernel were used to derive kernel features, and 50 latent components were used in the PLS model. Dynamic kernel features were not included, for simplicity. We used STRAIGHT to extract 24-dimensional Mel-Cepstral coefficients, 25 band aperiodicities (BAPs), and F_0.

VC-EVC: This is a many-to-many eigenvoice conversion (EVC) system [34]. The eigenvoice GMM (EV-GMM) was constructed from the training data from one pivot speaker in the ATR Japanese speech database [35], and 273 speakers (137 male, 136 female) from the JNAS database. [4] Settings were the same as in [36]. The 272-dimensional weight vectors were estimated by using the Part-A (small) training data. Covariance matrices in EV-GMM were not updated, i.e. the mean vectors of source and target speakers were independently updated. We used STRAIGHT to extract 24-dimensional Mel-Cepstral coefficients, 5 BAPs, and F_0. The number of mixture components was fixed at 128. The conversion method was applied only to the Mel-Cepstral coefficients.

VC-TVC: This is a tensor-based arbitrary voice conversion (TVC) system [36]. To construct the speaker space, the same Japanese dataset as in VC-EVC was used. The size of weight matrices which represent each speaker was set to 48 × 80. The same part of the **SAS** database and the same features as in VC-EVC were used, and again only the Mel-Cepstral coefficients were converted, without altering other features.

VC-FS: This is a frame selection voice conversion system, which is a simplified version of exemplar-based unit selection [37], using a single frame as an exemplar and without a concatenation (join) cost. We used the Part-A (small) data for training. The same features as in VC-KPLS were used, and once again only the Mel-Cepstral coefficients were converted.

VC-C1: As in VC-KPLS and VC-FS, STRAIGHT was used to extract Mel-Cepstral coefficients, BAPs and F_0. The first coefficient of the source speaker's Mel-Cepstral coefficients was converted by a linear transformation. This is the simplest voice conversion method, since it only changes the overall slope of the spectral envelope, and not any other speaker-specific features.

In all the voice conversion approaches, F_0 was converted by a global linear transformation: simple mean-variance normalisation. In VC-KPLS, VC-EVC, VC-TVC, VC-FS and VC-C1, source speaker BAPs were simply copied, without undergoing any conversion.

4. INITIAL BENCHMARKING EXPERIMENTS

To accompany the **SAS** database, we provide some benchmark speaker verification experimental results.

4.1. Speaker verification systems

We used two speaker verification systems representing the current state-of-the-art: Joint Factor Analysis (JFA) [38] and Probabilistic Linear Discriminant Analysis (PLDA) [39], under two enrolment scenarios, 5-utterance and 50-utterance. Both systems used the same front-end to extract acoustic features, comprising 19 dimension MFCC and energy features with delta and delta-delta coefficients. By excluding the static energy feature, 59-dimensional features were used in both systems. The AudioSeg toolkit was used to perform voice activity detection (VAD) [40]. In both systems, we used three Wall Street Journal (WSJ) databases (WSJ0, WSJ1, and WSJCAM) and the Resource Management database (RM1) for training the Universal Background Model (UBM) and the eigenspaces. From WSJ0 and WSJ1, only the SI training speakers were used. All speakers from the WSJCAM training, development and test sets were used. During scoring, T-norm was applied for both systems.

JFA: A Joint Factor Analysis system with a UBM of 512 components, and eigenvoice and eigenchannel spaces with 300 and 100 dimensions respectively. Cosine scoring was performed on the speaker variability vectors.

PLDA: Using the same UBM as in JFA, the PLDA approach operates in i-vector space, the dimension of which was set to 400. Because i-vectors have a heavy-tailed distribution, radial Gaussianization [41] was performed, then the i-vector dimension was reduced to 200 using linear discriminant analysis (LDA) and the within-class covariance matrices of the resulting vectors were whitened using

[4] http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html

Table 2. Initial spoofing results on the development and evaluation sets of **SAS** using the metrics of Equal Error Rate (EER) and False Alarm Rate (FAR) for the two variants (-5 and -50) of two speaker verification systems based on Joint Factor Analysis (JFA) or Probabilistic Linear Discriminant Analysis (PLDA).

		Development								Evaluation			
		EER				FAR				FAR			
	Spoofing	JFA-5	JFA-50	PLDA-5	PLDA-50	JFA-5	JFA-50	PLDA-5	PLDA-50	JFA-5	JFA-50	PLDA-5	PLDA-50
Male	NONE (Baseline)	3.29	1.29	1.44	0.66	3.29	1.29	1.44	0.66	3.43	1.40	1.44	0.66
	SS-SMALL	25.27	23.83	21.97	19.69	90.80	94.44	90.85	90.98	90.80	94.38	90.71	90.60
	SS-LARGE	27.47	25.95	23.96	22.15	93.59	97.23	94.11	94.46	93.64	97.32	93.68	94.05
	VC-FESTVOX	**30.09**	**30.36**	**28.94**	**27.97**	**95.55**	**98.32**	**98.60**	**99.20**	**95.46**	**98.44**	**98.41**	**99.11**
	VC-GMM	27.30	27.38	26.76	26.25	92.93	96.51	95.69	96.41	92.80	96.45	95.59	96.21
	VC-KPLS	19.60	18.24	20.96	20.11	76.76	84.56	89.45	89.51	77.10	84.70	89.19	89.46
	VC-TVC	19.32	17.69	20.03	18.94	73.40	80.32	84.73	84.45	73.68	80.67	84.46	84.37
	VC-EVC	15.64	13.12	16.20	14.73	62.34	67.67	80.12	78.83	62.68	67.94	80.09	78.92
	VC-FS	23.48	22.49	25.29	23.62	85.84	91.99	94.47	95.41	85.51	91.82	94.17	95.13
	VC-C1	3.60	1.44	1.69	0.86	4.48	2.23	2.28	1.25	4.66	2.16	2.24	1.15
Female	NONE (Baseline)	6.54	2.08	2.48	1.08	6.54	2.08	2.48	1.08	6.40	2.02	2.38	1.00
	SS-SMALL	23.76	17.90	19.49	17.78	79.03	77.01	83.53	89.48	79.43	77.53	83.96	89.88
	SS-LARGE	25.71	19.88	22.17	20.73	**83.39**	83.39	89.54	**94.23**	**83.58**	83.71	**89.90**	**94.55**
	VC-FESTVOX	**26.36**	**25.04**	**25.42**	**24.74**	82.06	**89.59**	**90.83**	93.20	82.45	**90.07**	88.69	91.27
	VC-GMM	26.32	24.84	23.95	23.65	81.32	88.38	88.70	91.88	81.88	89.02	89.37	92.41
	VC-KPLS	19.68	14.40	19.31	17.61	66.85	64.01	79.08	80.56	67.22	64.55	79.64	81.10
	VC-TVC	19.63	14.30	17.10	15.09	64.60	63.29	72.99	75.35	64.73	63.68	73.30	75.55
	VC-EVC	17.98	11.95	14.99	12.78	61.96	56.64	69.07	70.43	62.12	57.14	69.95	71.35
	VC-FS	20.89	15.94	21.08	19.70	68.71	71.19	81.82	87.51	69.12	71.52	82.27	87.78
	VC-C1	7.74	2.70	3.07	1.53	11.95	5.06	5.26	3.20	11.78	4.92	5.14	3.19

within-class covariance normalization (WCCN) [42]. The dimensionality of the resulting vectors was further reduced down to 100 by PLDA. Scoring was done with a likelihood ratio test.

In the two enrolment scenarios, the short enrolment utterances were merged into sessions of 5 before enrolment. Therefore, after merging, either 1 or 10 sessions were used in enrolment. For PLDA, in the 10 sessions case, i-vectors that were extracted from all 10 sessions were averaged, while for JFA, all features from all sessions were merged. We use JFA-5 and PLDA-5 to denote systems with 5 enrolment utterances (1 session), and JFA-50 and PLDA-50 for the 50-utterance (10 session) case.

4.2. Initial benchmarking results

We only report EERs and FARs for our initial speaker verification results, as the two measures are more related to spoofing. The results are presented in Table 2. Without surprise, the EERs and FARs for the baselines are very low, that is close or below 1% by JFA-50 and PLDA-50 systems, as the **SAS** database is clean without any channel or noise effects. However, the short duration of the trials prevents the EERs or FARs to go even lower.

Even through the ASV systems achieve very good speaker verification performance, they are extremely vulnerable to spoofing attacks. Even the most simple VC-C1 spoofing attack, which only changes the spectral slope of the source speaker, considerably increases the False Alarm Rate (FAR). The more sophisticated attacks using speech synthesis or voice conversion lead to FARs as high as 99.11%. In general, speech synthesis leads to FARs of over 90% for male and over 80 % for female, even for the SS-SMALL system which has access to only 24 utterances (Part-A) from the target speaker.

Voice conversion spoofing is sometimes an even more effective attack that speech synthesis. It is worth highlighting that the publicly-available voice conversion toolkit VC-FESTVOX is generally at least as effective as the other voice conversion and speech synthesis techniques. The second interesting observation is that although VC-EVC uses Japanese database to train eigenvoice for adaptation, it still increase FARs as high as other methods. Another observation is that even though more enrolment data is helpful to have lower EERs and FARs on non-spoofed data, it does not achieve lower error rates in the face of spoofing. These spoofing results are consistent with our previous findings on both telephone quality [11, 12] and clean speech [8, 9].

5. CONCLUSIONS

In this paper, we have introduced the first version of what we hope will become a standard dataset for spoofing and anti-spoofing research. Currently, the **SAS** corpus includes speech generated using nine spoofing methods, each of which comprises around 300000 spoofed trials. To set an initial benchmark, we have provided spoofing results when attacking two speaker verification systems. Without any countermeasures in place, these verification systems are extremely vulnerable to spoofing attacks from many of the nine spoofing methods included in **SAS** .

Our plan is to continue extending **SAS** by adding more sophisticated spoofing techniques, such as unit selection speech synthesis, frequency warping-based voice conversion, waveform modification-based voice conversion, phase-preserving voice conversion, and so on.

The current version of **SAS** is limited to text-independent speaker verification, so we plan to also create a database suitable for text-dependent speaker verification and include in that further spoofing attack methods such as replay spoofing.

In this paper, we only presented benchmark speaker verification results, to demonstrate the vulnerability of current systems to spoofing. In a future paper, we will present benchmark countermeasure results. There are also plans for challenge, in the spirit of NIST SRE evaluations or the Blizzard Challenge, to push forward research on spoofing and anti-spoofing countermeasures and to raise its visibility to the speaker verification, voice conversion, and speech synthesis communities.

Acknowledgements: this work was partially supported by EPSRC under Programme Grant EP/I031022/1 (Natural Speech Technology) and EP/J002526/1 (CAF), and by TUBITAK 1001 grant No 112E160.

6. REFERENCES

[1] Kong Aik Lee, Bin Ma, and Haizhou 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, J. Yamagishi, Z. Wu, F. Alegre, and P. DeLeon, "Speaker recognition anti-spoofing," in *Handbook of biometric anti-spoofing*, S. Marcel, S. Z. Li, and M. Nixon, Eds. Springer, 2014.

[3] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamgishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication (to appear)*, vol. 66, pp. 130–153, 2015.

[4] Zhizheng Wu and Haizhou Li, "Voice conversion versus speaker verification: an overview," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[5] Yee Wah Lau, Michael Wagner, and Dat Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. Int. Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004.

[6] R. Gonzalez 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.

[7] Zhizheng Wu, Sheng Gao, Eng Siong Chng, and Haizhou Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2014.

[8] Phillip L De Leon, Michael Pucher, and Junichi Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2010.

[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] Jean-François Bonastre, Driss Matrouf, and Corinne Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Interspeech*, 2007.

[11] Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, and Haizhou 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] Zhizheng Wu, Tomi Kinnunen, Eng Siong Chng, Haizhou Li, and Eliathamby 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] Federico Alegre, Ravichander Vipperla, Nicholas Evans, et al., "Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals," in *Proc. Interspeech*, 2012.

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

[15] Nicholas W D Evans, Junichi Yamagishi, and Tomi Kinnunen, "Spoofing and countermeasures for speaker verification: a need for standard corpora, protocols and metrics," in *IEEE Signal Processing Society Speech and language Technical Committee Newsletter*, 2013.

[16] Joseph P Campbell Jr, "Testing with the yoho cd-rom voice verification corpus," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1995.

[17] Federico Alegre, Asmaa Amehraye, and Nicholas Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *Proc. Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, 2013.

[18] Zhizheng Wu, Anthony Larcher, Kong Aik Lee, Eng Siong Chng, Tomi Kinnunen, and Haizhou Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints," in *Proc. Interspeech*, 2013.

[19] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[20] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[21] H. Zen, T. Nose, J. Yamagishi, S. Sako, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings of Sixth ISCA Workshop on Speech Synthesis*, Aug. 2007, pp. 294–299.

[22] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.B. Black, and T. Nose, *The HMM-based speech synthesis system (HTS) Version 2.2*, 2011, http://hts.sp.nitech.ac.jp/.

[23] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, pp. 187–207, 1999.

[24] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS system for Blizzard Challenge 2010," in *Proc. Blizzard Challenge 2010*, Kyoto, Japan, Sept. 2010.

[25] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[26] Christophe Veaux, Junichi Yamagishi, and Simon King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA*, 2013.

[27] Shinji Watanabe, Atsushi Nakamura, and Biing-Hwang(Fred) Juang, "Structural bayesian linear regression for hidden markov models," *Journal of Signal Processing Systems*, vol. 74, no. 3, pp. 341–358, 2014.

[28] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[29] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5-6, pp. 453–468, 1990.

[30] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for Mel-cepstral analysis of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 1992, pp. 137–140.

[31] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[32] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," in *Proc. Interspeech*, 2012.

[33] Elina Helander, Hanna Silén, Tuomas Virtanen, and Moncef Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[34] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Non-parallel training for many-to-many eigenvoice conversion," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4822–4825.

[35] A. Kurematsu, K. Takeda, Y. Sagisaka, H. Katagiri, S. Kuwabara, and K. Shikano, "Atr japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, pp. 357–363, 1990.

[36] D. Saito, N. Minematsu, and K. Hirose, "Effects of speaker adaptive training on tensor-based arbitrary speaker conversion," in *Proc. Interspeech*, 2012.

[37] Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li, "Exemplar-based unit selection for voice conversion utilizing temporal information," in *Proc. Interspeech*, 2013.

[38] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[39] Peng Li, Yun Fu, Umar Mohammed, James H. Elder, and Simon J.D. Prince, "Probabilistic models for inference about identity," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 144–157, January 2012.

[40] G. Gravier, M. Betser, and M. Ben, "audioseg: Audio segmentation toolkit, release 1.2," *IRISA*, January 2010.

[41] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011.

[42] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.