Anti-Spoofing for Text-Independent Speaker Verification: An Initial Database, Comparison of Countermeasures, and Human Performance

Zhizheng Wu*, Phillip L. De Leon, *Senior Member, IEEE,* Cenk Demiroglu, Ali Khodabakhsh, Simon King, *Fellow IEEE,* Zhen-Hua Ling, Daisuke Saito, Bryan Stewart, Tomoki Toda, Mirjam Wester, and Junichi Yamagishi, *Senior Member, IEEE*

***Abstract*—In this paper, we present a systematic study of the vulnerability of automatic speaker verification to a diverse range of spoofing attacks. We start with a thorough analysis of the spoofing effects of five speech synthesis and eight voice conversion systems, and the vulnerability of three speaker verification systems under those attacks. We then introduce a number of countermeasures to prevent spoofing attacks from both known and unknown attackers. Known attackers are spoofing systems whose output was used to train the countermeasures, whilst an unknown attacker is a spoofing system whose output was not available to the countermeasures during training. Finally, we benchmark automatic systems against human performance on both speaker verification and spoofing detection tasks.**

***Index Terms*—Speaker verification, speech synthesis, voice conversion, spoofing attack, anti-spoofing, countermeasure, security**

I. Introduction

The task of automatic speaker verification (ASV), sometimes described as a type of voice biometrics, is to accept or reject a claimed identity based on a speech sample. There are two types of ASV system: text-dependent and text-independent. Text-dependent ASV assumes constrained word content and is normally used in authentication applications because it can deliver the high accuracy required. However, text-independent ASV does not place constraints on word content, and is normally used in surveillance applications. For example, in call-center applications[1,2], a caller's identity can be verified during the course of a natural conversation without forcing the caller to speak a specific passphrase. Moreover, as such a verification process usually takes place under remote scenarios without any face-to-face contact, a *spoofing attack* – an attempt to manipulate a verification result by mimicking a target speaker's voice in person or by using computer-based techniques such as voice conversion or speech synthesis – is a fundamental concern. Hence, in this work, we focus on spoofing and anti-spoofing for text-independent ASV.

Due to a number of technical advances, notably channel and noise compensation techniques, ASV systems are being widely adopted in security applications [3], [4], [5], [6], [7]. A major concern, however, when deploying an ASV system, is its resilience to a spoofing attack. As identified in [8], there are at least four types of spoofing attack: impersonation [9], [10], [11], replay [12], [13], [14], speech synthesis [15], [16] and voice conversion [17], [18], [19], [20], [21]. Among the four types of spoofing attack, replay, speech synthesis, and voice conversion present the highest risk to ASV systems [8]. Although replay might be the most common spoofing technique which presents a risk to both text-dependent and text-independent ASV systems [12], [13], [14], it is not viable for the generation of utterances of specific content, such as would be required to maintain a live conversation in a call-center application. On the other hand, open-source software for state-of-the-art speech synthesis and voice conversion is readily available (e.g., Festival[3] and Festvox[4]), making these two approaches perhaps the most accessible and effective means to carry out spoofing attacks, and therefore presenting a serious risk to deployed ASV systems [8]. For that reason, the focus in this work is only on those two types of spoofing attacks.

A. Speech Synthesis and Voice Conversion Spoofing

Many studies have reported and analysed the vulnerability of ASV systems to speech synthesis and voice conversion spoofing. The potential vulnerability of ASV to synthetic

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. This article is an expanded version of [1], [2]

*Z. Wu is the correspondence author, and the remaining authors have been listed in alphabetical order to indicate equal contributions.

Z. Wu, S. King, M. Wester and J. Yamagishi are with the Centre for Speech Technology Research, University of Edinburgh, UK. e-mail: {zhizheng.wu, simon.king}@ed.ac.uk, {mwester, jyamagis}@inf.ed.ac.uk

P. L. De Leon and B. Stewart are with the Klipsch School of Electrical and Computer Engineering, New Mexico State University (NMSU), Las Cruces NM 88003 USA. e-mail: {pdeleon, brystewa}@nmsu.edu

A. Khodabakhsh and C. Demiroglu are with Ozyegin University, Turkey. e-mail: alikhodabakhsh@gmail.com, cenk.demiroglu@ozyegin.edu.tr

Z.-H. Ling is with University of Science and Technology of China, China. zhling@ustc.edu

D. Saito is with University of Tokyo, Japan. e-mail: dsk_saito@gavo.t.u-tokyo.ac.jp

T. Toda is with Information Technology Center, Nagoya University, Japan. e-mail:tomoki@icts.nagoya-u.ac.jp

[1] http://www.nuance.com/for-business/customer-service-solutions/voice-biometrics/freespeech/index.htm

[2] https://youtu.be/kyPTGoDyd_o

[3] http://www.cstr.ed.ac.uk/projects/festival/

[4] http://festvox.org/

speech was first evaluated in [22], [23]. An HMM-based speech synthesis system was used to spoof an HMM-based, text-prompted ASV system. They reported that the false acceptance rate (FAR) increased from 0% to over 70% under a speech synthesis spoofing attack. In [15], [16], the vulnerability of two ASV systems – a GMM-UBM system (Gaussian mixture models with a universal background model), and an SVM system (support vector machine using a GMM supervector) – was assessed using a speaker-adaptive, HMM-based speech synthesizer. Experiments using the Wall Street Journal (WSJ) corpus (283 speakers) [24] showed that FARs increased from 0.28% and 0.00% to 86% and 81% for GMM-UBM and SVM systems, respectively. These studies confirm the vulnerability of ASV systems to speech synthesis spoofing attack.

Voice conversion as a spoofing method has also been attracting increasing attention. The potential risk of voice conversion to a GMM ASV system was evaluated for the first time in [25], which used the YOHO database (138 speakers). In [26], [27], [17], text-independent GMM-UBM systems were assessed when faced with voice conversion spoofing on NIST speaker recognition evaluation (SRE) datasets. These studies showed an increase in FAR from around 10% to over 40% and confirmed the vulnerability of GMM-UBM systems to voice conversion spoofing attack.

Recent studies [18], [19] have evaluated more advanced ASV systems based on joint factor analysis (JFA), i-vectors, and probabilistic linear discriminative analysis (PLDA), on the NIST SRE 2006 database. The FARs of these systems increased five-fold from about 3% to over 17% under attacks from voice conversion spoofing.

B. Spoofing countermeasures

The vulnerability of ASV systems to spoofing attacks has led to the development of anti-spoofing techniques, often referred to as *countermeasures*. In [28], a synthetic speech detector based on the average inter-frame difference (AIFD) was proposed to discriminate between natural and synthetic speech. This countermeasure works well if the dynamic variation of the synthetic speech is different from that of natural speech; however, if global variance compensation is applied to the synthetic speech, the countermeasure becomes less effective [15].

In [29], [30], a synthetic speech detector based on image analysis of pitch-patterns was proposed for human versus synthetic speech discrimination. This countermeasure was based on the observation that there can be artefacts in the pitch contours generated by HMM-based speech synthesis. Experiments showed that features extracted from pitch-patterns can be used to significantly reduce the FAR for synthetic speech. The performance of the pitch-pattern countermeasure was not evaluated for detecting voice conversion spoofing.

In [31], a temporal modulation feature was proposed to detect synthetic speech generated by copy-synthesis. The modulation feature captures the long-term temporal distortion caused by independent frame-by-frame operations in speech synthesis. Experiments conducted on the WSJ database showed the effectiveness of the modulation feature when integrated with frame-based features. However, whether the detector is effective across a variety of speech synthesis and voice conversion spoofing attacks is unknown. Also using spectro-temporal information, a feature derived from local binary patterns [32] was employed to detect voice conversion and speech synthesis attacks in [33], [34].

Phase- and modified group delay-based features have also been proposed to detect voice conversion spoofing [35]. A cosine-normalised phase feature was derived from the phase spectrogram while the modified group delay feature contained both magnitude and phase information. Evaluation on the NIST SRE 2006 data confirmed the effectiveness of the proposed features. However, it remains unknown whether the phase-based features are also effective in detecting attacks from speech synthesisers using unknown vocoders. Another phase-based feature called the relative phase shift was proposed in [16], [36], [37] to detect speech synthesis spoofing, and was reported to achieve promising performance for vocoders using minimum phase rather than natural phase.

In [38], an average pair-wise distance (PWD) between consecutive feature vectors was employed to detect voice-converted speech, on the basis that the PWD feature is able to capture short-term variabilities, which might be lost during statistical averaging when generating converted speech. Although the PWD was shown to be effective against attacks from their own voice conversion system, this technique (which is similar to the AIFD feature proposed in [28]) might not be an effective countermeasure against systems that apply global variance enhancement.

In contrast to the above methods focusing on discriminative features, a probabilistic approach was proposed in [39], [40]. This approach uses the same front-end as ASV, but treats the synthetic speech as a signal passed through a synthesis filter. Experiments on the NIST SRE 2006 database showed comparable performance to feature-based countermeasures. In this work, we focus on feature-based anti-spoofing techniques, as they can be optimised independently without rebuilding the ASV systems.

C. Motivations and Contributions of this Work

In the literature, each study assumes a particular spoofing type (speech synthesis or voice conversion) and often just one variant (algorithm) of that type, then designs and evaluates a countermeasure for that specific, known attack. However, in practice it may not be possible to know the exact type of spoofing attack and therefore evaluations of ASV systems and countermeasures under a broad set of spoofing types are desirable. Most, if not all, previous studies have been unable to conduct a broader evaluation because of the lack of a standard, publicly-available spoofing database that contains a variety of spoofing attacks. To address this issue, we have previously developed a spoofing and anti-spoofing (**SAS**) database including both speech synthesis and voice conversion spoofing attacks [1]. This database includes spoofing speech from two different speech synthesis systems and seven different voice conversion systems.

Now, we first broaden the **SAS** database by including four more variants: three text-to-speech (TTS) synthesisers and one voice conversion system. They will be referred to as SS-SMALL-48, SS-LARGE-48, SS-MARY and VC-LSP[5], and are described in Section II.A.

We also develop a joint speaker verification and countermeasure evaluation protocol, then refine that evaluation protocol to enable better generalisability of countermeasures developed using the database. We include contributions from both the speech synthesis and speaker verification communities. This database is offered as a resource for researchers investigating generalised spoofing and anti-spoofing methods[6]. We hope that the availability of a standard database will contribute to reproducible research[7].

Second, with the **SAS** database, we conduct a comprehensive analysis of spoofing attacks on six different ASV systems. From this analysis we are able to determine which spoofing type and variant currently poses the greatest threat and how best to counter this threat. To the best of our knowledge, this study is the first evaluation of the vulnerability of ASV using such a diverse range of spoofing attacks and the most thorough analysis of the spoofing effects of speech synthesis and voice conversion spoofing systems under the same protocol.

Third, we present a comparison of several anti-spoofing countermeasures to discriminate between human and artificial speech. In our previous work, we applied cosine-normalised phase [35], modified group delay [35] and segment-based modulation features [31] to detect voice converted speech, and applied pitch pattern based features to detect synthetic speech [29], [30]. In this work, we evaluate these countermeasures against both spoofing types and propose to fuse decisions at the score level in order to leverage multiple, complementary sources of information to create stronger countermeasures. We also extend the segment-based modulation feature to an utterance-level feature, to account for long-term variations.

Finally, we perform listening tests to evaluate the ability of human listeners to discriminate between human and artificial speech[8]. Although the vulnerability of ASV systems in the face of spoofing attacks is known, some questions still remain unanswered. These include whether human perceptual ability is important in identifying spoofing and whether humans can achieve better performance than automatic approaches in detecting spoofing attacks. In this work, we attempt to answer these questions through a series of carefully-designed listening tests. In contrast to the human assisted speaker recognition (HASR) evaluation [43], we consider spoofing attacks in speaker verification and conduct listening tests for spoofing detection, which was not considered in the HASR evaluation.

[5]The four systems are new in this article while other systems have been published in a conference paper [1]. SS-SMALL-48 and SS-LARGE-48 allow us to analyse the effect of sampling rates of spoofing materials. SS-MARY is useful to understand the effect of waveform concatenation-based speech synthesis spoofing.

[6]Based on this database, a spoofing and countermeasure challenge [41], [42] has already been successfully organised as a special session of INTERSPEECH 2015.

[7]The SAS corpus is publicly available: http://dx.doi.org/10.7488/ds/252

[8]The preliminary version was published at INTERSPEECH 2015 [2] where we focused on human and automatic spoofing detection performance on wideband and narrowband data. The current work benchmarks automatic systems against human performance on speaker verification and spoofing detection tasks.

II. Database and Protocol

We extended our **SAS** database [1] by including additional artificial speech. The database is built from the freely available Voice Cloning Toolkit (VCTK) database of native speakers of British English[9]. 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 sentences are selected from newspapers, and the average duration of each sentence is about 2 seconds.

To design the spoofing database, we took speech data from VCTK comprising 45 male and 61 female speakers and divided each speaker's data into five parts:

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

In Parts B — E, sentences were randomly selected from newspapers without any repeating sentence across speakers. In Parts A and B, we have two versions, downsampled to 48 kHz and 16 kHz respectively, while in Parts C, D and E all signals are downsampled to 16 kHz. Parts A and B allow us to analyse the effects of sampling rate for spoofing attack. For training the spoofing systems, we designed two training sets. The small set consists of data only from Part A, while the large set comprises the data from Parts A and B together.

A. Spoofing systems

We implemented five speech synthesis (SS) and eight voice conversion (VC) spoofing systems, as summarised in Table I. These systems were built using both open-source software (to facilitate reproducible research) as well as our own state-of-the-art systems (to provide comprehensive results):

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-LARGE-16: An HMM-based TTS system built with the statistical parametric speech synthesis framework described in [44]. For speech analysis, the STRAIGHT vocoder with mixed excitation is used, which results in 60-dimensional Bark-Cepstral coefficients, $\log F_0$ and 25-dimensional band-limited aperiodicity measures [45], [46]. Speech data from 257 (115 male and 142 female) native speakers of British

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

TABLE I
SUMMARY OF THE SPOOFING SYSTEMS USED IN THIS PAPER. MGC, BAP AND F_0 MEAN MEL-GENERALISED CEPSTRAL (MGC) COEFFICIENTS, BAND APERIODICITY (BAP) AND FUNDAMENTAL FREQUENCY (F_0).

	Spoofing Algorithm	Sampling Rate	# training utterances	Vocoder	Features	Background data required?	Known or Unknown?	Open source Toolkit?
SS-LARGE-16	HMM TTS	16k	40	STRAIGHT	MGC, BAP, F_0	Yes	Known	Yes
SS-LARGE-48	HMM TTS	48k	40	STRAIGHT	MGC, BAP, F_0	Yes	Unknown	Yes
SS-SMALL-16	HMM TTS	16k	24	STRAIGHT	MGC, BAP, F_0	Yes	Known	Yes
SS-SMALL-48	HMM TTS	48k	24	STRAIGHT	MGC, BAP, F_0	Yes	Unknown	Yes
SS-MARY	Unit Selection TTS	16k	40	None	Waveform	No	Unknown	Yes
VC-C1	C1 VC	16k	24	STRAIGHT	MGC, BAP, F_0	No	Known	No
VC-EVC	Eigenvoice VC	16k	24	STRAIGHT	MGC, BAP, F_0	Yes	Unknown	No
VC-FEST	GMM VC	16k	24	MLSA	MGC, F_0	No	Known	Yes
VC-FS	Frame selection VC	16k	24	STRAIGHT	MGC, BAP, F_0	No	Known	No
VC-GMM	GMM VC	16k	24	STRAIGHT	MGC, BAP, F_0	No	Unknown	No
VC-KPLS	KPLS VC	16k	24	STRAIGHT	MGC, BAP, F_0	No	Unknown	No
VC-LSP	GMM VC	16k	24	STRAIGHT	LSP, F_0	No	Unknown	No
VC-TVC	Tensor VC	16k	24	STRAIGHT	MGC, BAP, F_0	Yes	Unknown	No

English is used to train the average voice model. In the speaker adaptation phase, the average voice model is transformed using structural variational Bayesian linear regression [47] followed by maximum a posteriori (MAP) adaptation, using the target speaker's data from Parts A and B. To synthesise speech, acoustic feature parameters are generated from the adapted HMMs using a parameter generation algorithm that considers global variance (GV) [48]. An excitation signal is generated using mixed excitation and pitch-synchronous overlap and add [49], and used to excite a Mel-logarithmic spectrum approximation (MLSA) filter [50] corresponding to the STRAIGHT Bark cepstrum, to create the final synthetic speech waveform.

SS-LARGE-48: Same as SS-LARGE-16, except that 48 kHz sample rate waveforms are used for adaptation. The use of 48 kHz data is motivated by findings in speech synthesis that speaker similarity can be improved significantly by using data at a higher sampling rate [51].

SS-SMALL-16: Same as SS-LARGE-16, except that only Part A of the target speaker data is used for adaptation.

SS-SMALL-48: Same as SS-SMALL-16, except that 48 kHz sample rate waveforms are used to adapt the average voice.

SS-MARY: Based on the Mary-TTS[10] unit selection synthesis system [52]. Waveform concatenation operates on diphone units. Candidate units for each position in the utterance are found using decision trees that query the linguistic features of the target diphone. A preselection algorithm is used to prune candidates that do not fit the context well. The target cost sums linguistic (target) and acoustic (join) costs. Candidate diphone and target diphone labels and their contexts are used to compute the linguistic sub-cost. Pitch and duration are used for the join cost. Dynamic programming is used to find the sequence of units with the minimum total target plus join cost. Concatenation takes place in the waveform domain, using pitch-synchronous overlap-add at unit boundaries.

VC-C1: The simplest voice conversion method, which modifies the spectral slope simply by shifting the first Mel-Generalised Cepstral coefficient (MGCs) [53]. No other speaker-specific features are changed. The STRAIGHT vocoder is used to extract MGCs, band aperiodicities (BAPs) and F_0.

VC-EVC: A many-to-many eigenvoice conversion (EVC) system [54]. The eigenvoice GMM (EV-GMM) is constructed from the training data of one pivot speaker in the ATR Japanese speech database [55], and 273 speakers (137 male, 136 female) from the JNAS database[11]. Settings are the same as in [56]. The 272-dimensional weight vectors are estimated by using the Part A of the training data. STRAIGHT is used to extract 24-dimensional MGCs, 5 BAPs, and F_0. The conversion function is applied only to the MGCs.

VC-FEST: The voice conversion toolkit provided by the open-source Festvox system. It is based on the algorithm proposed in [57], which is a joint density Gaussian mixture model with maximum likelihood parameter generation considering global variance. It is trained on the Part A set of parallel training data, keeping the default settings of the toolkit, except that the number of Gaussian components in the mixture distributions is set to 32.

VC-FS: A frame selection voice conversion system, which is a simplified version of exemplar-based unit selection [58], using a single frame as an exemplar and without a concatenation cost. We used the Part A set for training. The same features as in VC-C1 are used, and once again only the MGCs are converted.

VC-GMM: Another GMM-based voice conversion method very similar to VC-FEST but with some enhancements, which also uses the parallel training data from Part A. STRAIGHT is used to extract 24-dimensional MGCs, 5 BAPs, and F_0. The search range for F_0 extraction is automatically optimized speaker by speaker to reduce errors. Two GMMs are trained for separately converting the 1^{st} through 24^{th} MGCs and 5 BAPs. The number of mixture components is set to 32 for MGCs and 8 for BAPs, respectively. GV-based post-filtering [59] is used to enhance the variance of the converted spectral parameter trajectories.

VC-KPLS: Voice conversion using kernel partial least square (KPLS) regression [60], trained on the Part A parallel data. Three hundred reference vectors and a Gaussian kernel are used to derive kernel features and 50 latent components

[10]http://mary.dfki.de/

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

are used in the PLS model. Dynamic kernel features are not included, for simplicity. STRAIGHT is used to extract 24-dimensional MGCs, 25 BAPs, and F_0.

VC-TVC: Tensor-based arbitrary voice conversion (TVC) system [56]. To construct the speaker space, the same Japanese dataset as in VC-EVC is used. The size of the weight matrices that represent each speaker is set to 48 $\times$ 80. The same part of the **SAS** database and the same features as in VC-EVC are used, and again only MGCs are converted, without altering other features.

VC-LSP: This system is also based on the standard GMM-based voice conversion method similar to VC-GMM using the parallel training data from Part A. STRAIGHT is used as the speech analysis-synthesis method. 24-dimensional line spectral pairs (LSPs) and their delta coefficients are used as the spectral features. A 16-component GMM is trained for the modelling of joint LSP feature vectors. For each component, the four blocks of its covariance matrix are set to be diagonal. No quality enhancement or post-filtering techniques are applied during the reconstruction of converted speech.

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

B. Speaker Verification and Countermeasure Evaluation Protocol

For the evaluation of ASV systems, enrolment data for each client (speaker) were selected from Part C under two conditions: 5-utterance or 50-utterance enrolments. For five utterances, this is about 5-10 seconds of speech while for 50 utterances it is about 1 minute of speech. The development set, used to tune the ASV system and decide thresholds, was taken from Part D and involves both genuine and impostor trials. All utterances from a client speaker in Part D were used as genuine trials, and this results in 1498 male and 1999 female genuine trials. For the impostor trials, ten 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 12981 male and 17462 female impostor trials. The evaluation set is taken from Part E and is arranged into genuine and imposter trials in a similar way to the development set, with 4053 male and 5351 female genuine trials, and 32833 male and 46736 female impostor trials. A summary of the development and evaluation sets is shown in Table II.

TABLE II
NUMBER OF TRIALS IN THE DEVELOPMENT AND EVALUATION SETS.

	Development		Evaluation	
	Male	Female	Male	Female
Target speakers	15	20	20	26
Genuine trials	1498	1999	4053	5351
Impostor trials	12981	17462	32833	46736
Spoofed trials	12981×5	17462×5	32833×13	46736×13

We used the synthetic speech and voice conversion systems described above to generate artificial speech for both development and evaluation sets. During the execution of spoofing attacks, the transcript of an impostor trial was used as the textual input to each speech synthesis system, and the speech signal of the impostor trial was the input to each voice conversion 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). In this way, the number of spoofed trials of one spoofing system is exactly the same as the number of impostor trials presented in Table II. This allows a fair comparison between non-spoofed and spoofed speaker verification results. Only five of the available spoofing systems were used during development, with all thirteen spoofing systems (Table I) being run on the evaluation set. Hence, the number of total spoofed trials is 12981×5 and 17462×5 for males and females, respectively, for the development set, and 32833×13 and 46736×13 for male and female speakers, respectively, for the evaluation set.

TABLE III
NUMBER OF SPEAKERS AND TRIALS FOR TRAINING, DEVELOPMENT AND EVALUATION SETS OF THE COUNTERMEASURE PROTOCOL.

	#Speakers		#Trials	
	Male	Female	Genuine	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	152215
Evaluation	20	26	9404	1034397

In the countermeasure evaluation protocol, we used a further 25 speakers' voices as training data and only implemented five attacks (as known attacks) on the training set. The 25 speakers do not appear in the development and evaluation sets for ASV, and this allows us to develop speaker- and gender-independent countermeasures. For countermeasure development and evaluation sets, the same speakers and same spoofed trials are used as those for ASV. This allows us to integrate countermeasures with ASV systems and to evaluate the integration performance. A summary of the countermeasure protocol is presented in Table III.

III. SPEAKER VERIFICATION SYSTEMS

We used three classical ASV systems: Gaussian Mixture Models with a Universal Background Model (GMM-UBM) [61], Joint Factor Analysis (JFA) [62] and i-vector with Probabilistic Linear Discriminant Analysis (PLDA) [63]. In this paper, we use **PLDA** to refer to this i-vector-PLDA system. Each system was implemented under the two enrolment scenarios: 5-utterance and 50-utterance enrolment. All systems used the same front-end to extract acoustic features: 19-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) plus log-energy with delta and delta-delta coefficients. By excluding the static energy feature (but retaining its delta and delta-delta), 59-dimensional feature vectors are obtained. To extract MFCCs, we applied a Hamming analysis window, the size of which is 25 ms with a 10-ms shift, and we employed a mel-filter bank with 24 channels. We note that C0 is not

TABLE IV
STATISTICS OF WALL STREET JOURNAL (WSJ0, WSJ0, WSJCAM) AND RESOURCE MANAGEMENT (RM) DATABASES USED TO TRAIN UBM AND EIGENSPACES.

	WSJ0+WSJ1		RM		WSJCAM		Total	
	Male	Female	Male	Female	Male	Female	Male	Female
Speakers	149	152	108	52	76	59	333	264
Utterances	14900	15199	7881	3982	6697	5148	29478	24329
Hours	∼26.3	∼27.9	∼5.6	∼2.9	∼12.1	∼9.5	∼44	∼40.4

retained in the extracted MFCCs. In practice, the SPro toolkit[12] was used to extract MFCCs. The AudioSeg toolkit was used to perform voice activity detection (VAD) [64].

GMM-UBM: with 512 Gaussian components in the UBM, and a client speaker model obtained by performing maximum *a posteriori* (MAP) adaptation, with the relevance factor set to 10. Only mean vectors were adapted, keeping diagonal covariance matrices and mixture weights the same as in the UBM.

JFA: using a UBM with the same 512 components as the GMM-UBM as well as eigenvoice and eigenchannel spaces with 300 and 100 dimensions, respectively. Cosine scoring was performed on the speaker variability vectors.

PLDA: a PLDA system operating in i-vector space. An i-vector is a low-dimensional vector to represent a speaker- and channel-dependent GMM supervector M through a low rank matrix T, as $M = m + Tw$, where m is a speaker- and channel-independent supervector, which is realised by a UBM supervector in this work; T is also called the total variability matrix; and w is the i-vector. In this work, 400-dimensional i-vectors were extracted with the maximum a posteriori (MAP) criterion and using the same UBM as the JFA system. Linear discriminant analysis (LDA) was first applied to reduce the i-vector dimension to 200. Then, i-vectors were centred, length-normalised, and whitened. The whitening transformation was learned from i-vectors in the development set. After that, a Gaussian PLDA model was trained using the expectation-maximisation (EM) algorithm which was run for 20 iterations. The rank of the eigenspace (number of columns in the eigenmatrix) was set to 100. Scoring was done with a log-likelihood ratio test. In practice, the MSR Identity Toolbox [65] was used to implement the PLDA system.

We used three WSJ databases (WSJ0, WSJ1, and WSJCAM) and the Resource Management database (RM1) for training the UBM, eigenspaces, and LDA. The statistics of the three databases are presented in Table IV. The sampling rate of all four database is 16 kHz. We note that our preliminary experimental results suggested that WSJCAM was very useful for improving verification performance. The maximum likelihood criterion was employed to train the UBM and eigenspaces while the Fisher criterion was used to train LDA.

The 50 enrolment utterances were merged into 10 sessions (each being the concatenation of 5 utterances); either 1 or 10 of these sessions were used in enrolment, for the two enrolment scenarios. For PLDA, when using 10 enrolment sessions, i-vectors were extracted from each session then averaged as suggested in [66]; for JFA, all features from all sessions were merged. We denote the ASV systems with 5 enrolment utterances (presented as 1 session) as GMM-UBM-5, JFA-5 or PLDA-5 and those with 50 enrolment utterances (presented as 10 sessions) as GMM-UBM-50, JFA-50 or PLDA-50.

[12]Available at: http://www.irisa.fr/metiss/guig/spro/

IV. ANTI-SPOOFING COUNTERMEASURES

We now examine five countermeasures[13], described below along with the features they are based on, and then propose a fusion of these countermeasures in order to learn complementary information and improve anti-spoofing performance.

Given a speech signal $x(n)$, short-time Fourier analysis can be applied to transform the signal from the time domain to the frequency domain by assuming the signal is quasi-stationary within a short time frame, e.g., 25ms. The short-time Fourier transform of the speech signal can be represented as follows:

$$X(\omega) = |X(\omega)|e^{j\varphi(\omega)}, \tag{1}$$

where $X(\omega)$ is the complex spectrum, $|X(\omega)|$ is the magnitude spectrum and $\varphi(\omega)$ is the phase spectrum. It is usual to define $|X(\omega)|^2$ as the power spectrum, from which features that only contain magnitude information, e.g., MFCCs, can be derived. The proposed feature-based countermeasures are derived from the complex spectrum $X(\omega)$ that has two parts: a real part $X_R(\omega)$ and an imaginary part $X_I(\omega)$, and from which the phase spectrum $\varphi(\omega)$ can be obtained.

To extract frame-wise features, we employ a hamming window, the size of which is 25ms, with a 5ms shift. The FFT length is set to 512.

A. Cosine Normalised Phase Feature

Even though phase information is important in human speech perception [67], most speech synthesis and voice conversion systems use a simplified, minimum phase model which may introduce artefacts into the phase spectrum. The cosine normalised phase (CosPh) feature is derived from the phase spectrum, and can be used to discriminate between human and synthetic speech. The feature is computed as follows:

1) Unwrap the phase spectrum.
2) Compute the CosPh spectrum by applying the cosine function to the spectrum in 1) to normalise to $[-1.0,\ 1.0]$.
3) Apply a discrete cosine transform (DCT) to the spectrum in 2).

[13]The cosine normalised phase feature, modified group delay cepstral feature, segment-based modulation feature and pitch pattern feature based countermeasures have been presented in our previous conference papers [35], [31], [30]. The current study examines the generalisation abilities of each individual countermeasure and their combination in the face of various spoofing attacks.

4) Keep the first 18 cepstral coeffcients, and compute their delta and delta-delta coeffcients as features.

By normalizing the values of the unwrapped phase spectrum, we can simplify subsequent statistical modeling. We note that the motivation for applying the DCT is decorrelation and dimensionality reduction; C0 is not retained.

B. Modified Group Delay Cepstral Feature

In addition to the artefacts in the phase spectrum, the statistical averaging inherent in parametric modeling of the magnitude spectrum may also introduce artefacts, such as oversmoothed spectral envelopes. The use of both phase and magnitude spectra can therefore be useful for detecting synthetic speech. The Modified Group Delay Cepstral Coefficients (MGDCCs) can be used to detect artefacts in both spectra of synthetic speech. The MGDCC feature has also been used in speech recognition [68] and speaker verification [69]. The MGDCCs are derived from the complex spectrum as follows:

1) Apply the fast Fourier transform (FFT) to a windowed speech signal, $x(n)$ and $nx(n)$ to compute $X(\omega)$ and $Y(\omega)$, respectively. Here $nx(n)$ is the re-scaled signal of $x(n)$.
2) Compute the cepstrally-smoothed power spectrum[14] $|S(\omega)|^2$ of $|X(\omega)|^2$.
3) Compute the MGD spectrum (R and I denote the real and imaginary parts of the spectrum)

$$\tau_\rho(w) = \frac{X_R(w)Y_R(w) + Y_I(w)X_I(w)}{|S(w)|^{2\rho}}. \quad (2)$$

4) Reshape $\tau_\rho(w)$ as

$$\tau_{\rho,\gamma}(w) = \frac{\tau_\rho(w)}{|\tau_\rho(w)|}|\tau_\rho(w)|^\gamma. \quad (3)$$

5) Apply the DCT to $\tau_{\rho,\gamma}(w)$ and keep the first 18 cepstral coefficients with their delta and delta-delta coefficients as MGDCC features.

In (2) and (3), ρ and γ are two weighting variables that control the shape of the MGD spectrum. We set $\rho = 0.7$ and $\gamma = 0.2$ based on the performance on the development set.

C. Segment-Based Modulation Feature

In speech synthesis and voice conversion, the speech signal is usually divided into overlapping frames for modeling, and this frame-by-frame or state-by-state modeling may introduce artefacts in the temporal domain due to the independence assumptions made by the underlying statistical model. These temporal artefacts are evident in the modulation domain and can be used to detect synthetic and voice-converted speech. The Segment-based Modulation Feature (SMF) is extracted from the MGD cepstrogram based on our previous work [31]. The procedure for computing the SMF is illustrated in Fig. 1 and described as follows:

[14]Cepstrally-smoothed spectrum is obtained through the following steps: a) compute the log-amplitude spectrum from $X(\omega)$, and apply a median filter to smooth the spectrum; b) apply the DCT to the log spectrum and keep the first 30 cepstral coefficients; c) apply the inverse DCT to the cepstral coeffcients to obtain the cepstrally-smoothed spectrum $S(\omega)$.

1) Divide the 18-dimensional MGD spectrogram into overlapping segments using a 50-frame window with 20-frame shift.
2) Apply mean and variance normalisation to the MGD trajectory of each dimension to make it have zero mean and unit variance[15].
3) Take the FFT of the normalised 18-dimensional trajectories to compute modulation spectra.
4) Concatenate the modulation spectra in one cepstrogram segment into a supervector, and use this as the SMF feature vector.
5) Average all the SMF vectors of one utterance to get an average feature vector. This averaged feature vector will be used as the feature vector for the utterance.

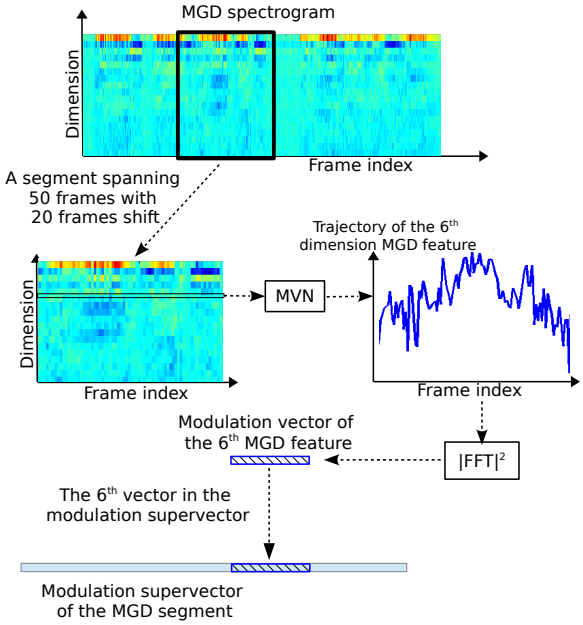

Fig. 1. The process to extract Segment-based Modulation Features (SMFs) from modified group delay cepstral features.

In practice, we used a 64-point FFT to extract a 32-dimensional modulation spectrum for each MGD trajectory. Hence, the modulation supervector of each segment is $18 \times 32 = 576$. We pass this supervector to a support vector machine (SVM) for classification. In practice, we employed the LIBSVM toolkit [70] to implement the SVM. We used a radial basis kernel, and set the penalty factor to 34.

D. Utterance-Based Modulation Feature

To extract the segment-based modulation feature, a speech signal needs to be divided into short segments first and then the corresponding modulation features are extracted for each segment. An alternative approach is to extract modulation features at the utterance level, to obtain Utterance-based Modulation Features (UMFs).

The process to extract UMFs is similar to that of SMFs, but only steps 2 – 4 are applied, without dividing the utterances

[15]The motivation to perform mean-variance normalisation is to make the trajectory of each dimension in the same scale.

into frames. In practice, we used a 1024-point FFT to extract the modulation spectrum for each MGD trajectory, then applied a DCT to the modulation spectrum, and after that kept the first 32 coefficients as features. Hence, the dimensionalities of UMF and SMF for each utterance are the same: 576. Again, we pass the feature vector to an SVM for classification. The configuration of the SVM here is the same as that for SMF in Section IV-C.

E. Pitch Pattern Feature

The prosody of synthetic speech is generally not the same as natural speech [71] and therefore the pitch pattern is another good candidate feature for a countermeasure. The pitch pattern, $\phi[n,m]$, is calculated by dividing the short-range autocorrelation function, $r[n,m]$ by a normalization function, $p[n,m]$ which is proportional to the frame energy [72]

$$\phi[n,m] = \frac{r[n,m]}{p[n,m]} \tag{4}$$

where

$$r[n,m] = \sum_{k=-m/2}^{m/2} x[n+k-m/2]x[n+k+m/2], \tag{5}$$

$$p[n,m] = \frac{1}{2}\sum_{k=-m/2}^{m/2} x^2[n+k-m/2] + \frac{1}{2}\sum_{k=-m/2}^{m/2} x^2[n+k+m/2], \tag{6}$$

and n, m are the sample instant and lag, respectively, over which the autocorrelation is computed. The lag parameter is chosen such that pitch frequencies can be observed [72]; in this work, we choose $32 \leq m \leq 320$ for a sample rate of 16kHz.

Once the pitch pattern is computed, we segment it into a binary pitch pattern image through the rule

$$\phi_{\text{seg}}[n,m] = \begin{cases} 1, & \phi[n,m] \geq \theta \\ 0, & \phi[n,m] < \theta \end{cases} \tag{7}$$

where θ is a threshold; we set $\theta = 1/\sqrt{2}$ for all n, based on preliminary results on the development set. An example pitch pattern image is shown in Fig. 2.

Extracting features from the pitch pattern is a two-step process: 1) computation of the pitch pattern; 2) image analysis. First, the pitch pattern is computed using (4) and segmented using (7) to form a binary image. In the second step, image processing of the segmented binary pitch pattern is performed in order to extract the connected components (CCs), i.e., black regions in Fig. 2. This processing includes determining the bounding box and area of a CC, which are then used to distinguish between two types of CC: pitch pattern connected components (PPCC) and irregularly-shaped components or artefacts.

The resulting CCs are then analysed and the mean pitch stability μ_s, mean pitch stability range μ_R, and time support (TS) of each CC are computed as in [29]. The proposed image processing-based approach determines parameters on a per-connected component basis and then computes statistics over the connected components of the utterance. The six element utterance feature vector used for classification contains μ_R and the TS of the artefacts, the number of artefacts, μ_S and TS of the PPCC, and standard deviation of the TS of PPCC. Other utterance features were considered during the training and development stage but were found not to contribute to the classifier accuracy.

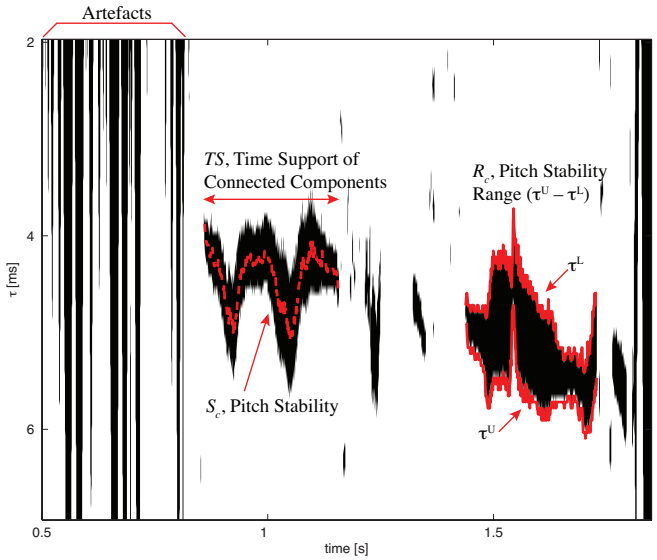

Fig. 2. Example binary pitch pattern image illustrating pitch stability S_c, pitch stability range R_c, upper edge τ^{U}, lower edge τ^{L}, connected component time support, and artefacts.

For the pitch pattern countermeasure, a maximum likelihood classifier based on the log-likelihoods computed from the utterance feature vectors was used for classification. During training, human and spoofing utterance feature vectors were modeled as multivariate Gaussian distributions with full covariance matrices. During testing, the utterance is determined to be human if the log-likelihood ratio is greater than a threshold calibrated to produce equal error rate (EER) on the development set.

F. Fused countermeasure

To benefit from the multiple feature-based countermeasures, we propose a fused countermeasure. In speaker verification, system fusion is one way to combine multiple individual speaker verification systems to achieve better performance [73], [74]. A similar strategy can be applied for anti-spoofing, as each feature-based countermeasure discussed above has its own pros and cons. For example, the pitch pattern feature-based countermeasure is expected to work well in detecting waveform concatenation based spoofing attacks, while other countermeasures are expected to detect phase and temporal artefacts. It is expected the fused countermeasure can benefit from the pros of each individual countermeasure.

We perform linear fusion at the score level. We first train a linear fusion function on the development set which only contains known attacks, and then apply the fusion function

on the evaluation scores; finally, the fused score is used to discriminate between human and spoofed speech. In practice, we used the BOSARIS Toolkit[16] to train the fusion function.

V. Experiments

A. Evaluation Metric

In both speaker verification and spoofing detection, there are two types of errors: 1) genuine or human speech is accepted as impostor or spoofed speech; 2) impostor or spoofed speech is accepted as genuine or human. The first type of error is a false rejection error, while the second type is a false acceptance. When the false acceptance rate (FAR) equals to the false rejection rate (FRR), we are at the equal error rate (EER) point. In this work, when reporting the false acceptance rates (FARs) and the false rejection rates (FRRs) for a specific spoofing algorithm, the decision threshold is set to achieve the EER operating point for that spoofing algorithm. When reporting overall spoofing performance, all the spoofed samples are pooled together and treated as one (unknown) spoofing algorithm when setting the threshold, because in practice one may not know the exact type of spoofing algorithm.

B. Spoofing ASV Systems without Countermeasures

We evaluated the performance of the ASV systems for the various synthetic speech and voice conversion variants described in Section II-A. Prior to the evaluation, the ASV decision threshold was set to the EER point on the development set, using only human speech.

Speaker verification results are presented in Table V. The FARs for the baseline experiment, which uses only human speech, are low (as expected) because the SAS database has near-ideal recordings, free from channel and background noise. In particular, the lowest FARs for GMM, JFA and PLDA systems are 0.09%, 1.25% and 1.16%, respectively. Note that the short duration of the trials preculdes even lower FARs and FRRs.

Whilst the ASV systems achieve excellent verification performance, they are still vulnerable to spoofing. The simple VC-C1 spoofing attack, which only modifies the spectral slope of the source speaker, increases FAR for nearly every ASV system. The attacks using speech synthesis or voice conversion, with more advanced algorithms, lead to FARs as high as 99.95%. On average, speech synthesis leads to FARs of over 95% for male voices and over 80% for female voices, and more sophisticated voice conversion algorithms lead to FARs of close to 80% for both male and female voices. These observations are consistent with previous studies on clean speech [16] and telephone quality speech [18], [19], and confirm the vulnerability of ASV systems to a diverse range of spoofing attacks. In general, our experiments suggest that it is easier to spoof male speakers than female speakers in the sense that the FARs for the various spoofing attacks for female speakers are generally lower than that for male speakers. We speculate that it is relatively harder to model female speech or perform female-to-female conversion due to the higher variability of female speech.

Although ASV systems that have more enrolment data available to them give lower FARs in the baseline case, they are not necessarily more resistant to spoofing attack. For example, under the VC-FEST attack, the FARs of JFA-5 and PLDA-5 male systems are 91.25% and 97.41%, respectively, and the FARs of JFA-50 and PLDA-50 are even higher at 97.71% and 99.54%, respectively. Similar patterns can be observed for other spoofing algorithms, as well as for female speech.

From the perspective of spoofing, the first interesting observation is that voice conversion is as effective at spoofing as speech synthesis, given the same amount of training data. Most of the speech synthesis systems used in this work require a large amount of data to train the average voice model, which is adapted to the target. On the other hand, most voice conversion algorithms, including VC-FEST, VC-GMM and VC-FS, only need source and target speech data to train their conversion functions. Voice conversion spoofing is sometimes even more effective than speech synthesis. It is worth highlighting that the publicly-available voice conversion toolkit VC-FEST is at least as effective as the other voice conversion and speech synthesis techniques.

The second interesting observation is that, although VC-TVC and VC-EVC use a Japanese database to train eigenvoices for adaptation to English data, these methods still increase FARs as much as the other variants. This suggests that attackers could use alternate speech resources, i.e. speech corpora in another language, if they cannot find enough resources for the target language.

The third observation is that the use of higher sampling rate training data in speech synthesis results in higher FARs of ASV systems. This suggests that such data includes more speaker-specific characteristics and that attackers can use this to conduct more effective spoofing if they have access to such data.

The last observation is that more training data can improve the effectiveness of speech synthesis and voice conversion spoofing systems. Comparing SS-SMALL-16k and SS-LARGE-16k, using 40 instead of 24 training utterances results in an increase of about 4% absolute FAR. In contrast, using more enrollment data for ASV systems does not seem to be helpful in defending against spoofing attacks (except VC-C1), although it does improve the baseline ASV performance without spoofing. We speculate that, as the spoofed speech sounds more like the target speaker, it will achieve higher likelihood scores under any target speaker model that has been trained using more enrollment data, and hence results in higher FARs. This also explains why ASV systems with more enrollment data succeed in defending against the VC-C1 attack, which can be easily distinguished by the human ear in terms of speaker similarity, as shown in Table VIII.

Given the wide-ranging spoofing results in Table V and the above observations, it is clear that countermeasures are needed. So, we next present an evaluation and analysis of a range of countermeasures, including a proposed new fused countermeasure.

[16]https://sites.google.com/site/bosaristoolkit/

TABLE V
FALSE ACCEPTANCE RATES (FARS) IN %, ON THE EVALUATION SET FOR THE TWO VARIANTS (-5 AND -50) OF THREE SPEAKER VERIFICATION SYSTEMS BASED ON: A GAUSSIAN MIXTURE MODEL WITH UNIVERSAL BACKGROUND MODEL (GMM-UBM); JOINT FACTOR ANALYSIS (JFA); AND PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS (PLDA). THE DECISION THRESHOLD IS SET TO THE EQUAL ERROR RATE (EER) POINT ON THE DEVELOPMENT SET.

	Male						Female					
Spoofing	GMM-UBM-5	GMM-UBM-50	JFA-5	JFA-50	PLDA-5	PLDA-50	GMM-UBM-5	GMM-UBM-50	JFA-5	JFA-50	PLDA-5	PLDA-50
Baseline	4.05	0.09	2.76	1.25	1.41	1.16	11.10	0.66	6.24	2.47	1.52	0.99
SS-LARGE-16	79.86	97.86	88.62	96.17	93.45	97.76	90.13	89.34	84.31	84.65	86.04	95.95
SS-LARGE-48	97.35	99.95	97.62	98.93	99.12	99.09	98.52	99.28	90.58	94.28	94.80	98.39
SS-MARY	86.57	99.39	91.09	96.81	96.77	98.74	95.23	99.17	91.37	95.11	95.28	98.10
SS-SMALL-16	75.65	91.62	83.64	91.25	89.21	94.87	86.49	81.72	80.60	77.97	81.60	93.14
SS-SMALL-48	95.63	98.89	94.97	95.75	97.07	96.63	97.44	97.63	86.46	90.36	93.02	96.86
VC-C1	4.78	0.11	2.62	1.46	1.83	1.67	17.68	1.94	12.71	6.80	3.92	3.56
VC-EVC	50.64	56.63	43.38	58.52	69.84	79.50	71.60	67.45	67.82	66.50	72.14	79.10
VC-FEST	79.67	98.29	91.25	97.71	97.41	99.54	91.30	94.39	85.77	91.76	86.11	93.53
VC-FS	79.12	98.65	78.68	91.62	91.05	96.16	86.61	94.77	71.19	75.32	79.37	90.33
VC-GMM	76.03	97.35	89.14	96.22	95.10	98.70	91.94	97.53	85.72	92.93	90.57	97.42
VC-KPLS	59.60	72.76	61.92	82.90	81.17	89.31	77.30	72.96	70.99	71.72	80.87	86.32
VC-LSP	57.98	71.57	51.71	68.37	74.99	89.82	75.26	75.44	65.30	60.27	70.99	75.14
VC-TVC	58.64	70.94	63.28	78.75	80.20	87.14	77.16	75.37	70.87	71.41	74.83	82.20

C. Evaluation of Stand-Alone Countermeasures

We conducted experiments to evaluate the performance of stand-alone countermeasures, i.e. their ability to discriminate between human and artificial speech. When training countermeasures, five of the spoofing systems listed in Table I, were used: SS-SMALL-16, SS-LARGE-16, VC-C1, VC-FEST and VC-FS.

For MFC, CosPh, MGD and PP features, GMM-based maximum likelihood classifiers were employed, while for SMS and UMS features, SVM classifiers were used. Whilst many combinations of features and classifier could of course be imagined, these choices give us a representative range of countermeasures to compare. For each countermeasure, the detection threshold was set to achieve the EER point on the development set under all five known attacks, and then the countermeasure was applied to the evaluation set to compute the FARs shown in Table VI. These results show that the frame-based features MFCC, CosPh and MGD achieve better performance than the long-term features SMS, UMS and PP. Even though the modulation features SMS and UMS are derived from the MGD features, they do not perform as well as frame-based MGD features. This observation is consistent with our previous work [31]. In the database, due to the short duration of trials, long-term features generally only provide a rather small number of feature vectors per utterance.

In respect of the frame-based features, the MGD-based countermeasure achieves the best overall performance in terms of low FARs and works well at detecting most types of spoofed speech with the notable exception of the SS-MARY attack. The MGD features include both magnitude and phase spectrum information, whereas MFCCs only capture magnitude spectrum and CosPh only phase spectrum. With respect to long-term features, both SMS and UMS perform well at detecting statistical parametric speech synthesis spoofing, yet fail to detect most of the voice conversion algorithms or unit selection speech synthesis.

The pitch pattern countermeasure detects synthetic speech well, but does not detect some voice conversion speech such as that from VC-C1, VC-FEST, VC-KPLS and VC-LSP. This is probably due to the fact that speech synthesis usually predicts fundamental frequency (F0) from text (and so produces rather unnatural trajectories) whereas voice conversion usually copies a source speaker's F0 trajectories to generate a target speaker's voice. Hence, voice conversion introduces fewer pitch pattern artefacts than speech synthesis. We note that the pitch pattern countermeasure achieves the best performance of 1.96% FAR against the SS-MARY unit selection synthesis attack.

TABLE VI
SPOOFING COUNTERMEASURE RESULTS IN TERMS OF FALSE ACCEPTANCE RATE (FAR) IN % ON THE EVALUATION SET. THE DECISION THRESHOLD IS SET TO THE EER POINT ON THE DEVELOPMENT SET. THE FIRST GROUP OF 5 ATTACK METHODS IS KNOWN AND THE REMAINING 8 ARE UNKNOWN.

	MFC	CosPh	MGD	SMS	UMS	PP	Fusion
SS-SMALL-16	0.01	1.41	0.10	5.31	8.43	0.00	0.00
SS-LARGE-16	0.01	1.03	0.11	5.44	8.10	0.00	0.00
VC-C1	27.08	0.44	4.07	45.20	33.70	68.73	0.80
VC-FEST	0.84	21.89	4.61	37.76	39.75	60.56	0.43
VC-FS	0.10	0.04	0.07	4.18	4.80	7.66	0.00
SS-LARGE-48	0.01	0.00	0.00	0.62	0.46	0.00	0.00
SS-MARY	89.30	92.76	93.92	81.81	87.91	1.96	97.76
SS-SMALL-48	0.00	0.01	0.00	0.71	0.43	0.00	0.00
VC-EVC	2.72	0.01	1.87	23.28	4.18	0.00	0.02
VC-GMM	1.61	19.68	4.37	37.93	33.08	9.86	0.79
VC-KPLS	1.15	0.06	0.54	21.08	7.56	68.64	0.00
VC-LSP	4.86	0.03	0.84	56.93	19.46	73.26	0.15
VC-TVC	2.97	0.02	1.58	23.11	5.31	0.01	0.01
known	5.61	4.96	1.79	19.58	18.95	27.39	0.25
unknown	12.83	14.07	12.89	30.68	19.80	19.22	12.34
all attacks	10.05	10.57	8.62	26.41	19.47	22.36	7.69

In general, most of the countermeasures achieve better performance for known attacks than for unknown attacks, as spoofing data from known attacks are available for training countermeasures and those from unknown attacks are not available to train the detectors. From the perspective of spoofing algorithms, SS-MARY is the most difficult to detect, and this is presumed to be due to the fact that it uses original waveforms to generate spoofed speech and thus introduces fewer artefacts when compared with other methods.

We also fused the six individual countermeasures at the score level to create a new countermeasure as detailed in Section IV-F. The linear combination weights for MFC, CosPh, MGD, SMS, UMS and PP countermeasures are 26.71, 9.56,

TABLE VII
FALSE ACCEPTANCE RATES (FARS) ON THE EVALUATION SET FOR THE TWO VARIANTS (5 AND 50) OF THREE SPEAKER VERIFICATION SYSTEMS WITH INTEGRATED COUNTERMEASURE. THESE ASV SYSTEMS ARE EACH BASED ON A GAUSSIAN MIXTURE MODEL WITH UNIVERSAL BACKGROUND MODEL (GMM-UBM), JOINT FACTOR ANALYSIS (JFA) OR PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS (PLDA). THE DECISION THRESHOLD IS SET TO THE ASV EQUAL ERROR RATE (EER) POINT ON THE DEVELOPMENT SET USING ONLY HUMAN SPEECH.

	Male						Female					
Spoofing	GMM-UBM-5	GMM-UBM-50	JFA-5	JFA-50	PLDA-5	PLDA-50	GMM-UBM-5	GMM-UBM-50	JFA-5	JFA-50	PLDA-5	PLDA-50
SS-LARGE-16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SS-LARGE-48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SS-MARY	84.91	97.48	89.46	95.01	95.08	96.91	94.27	98.09	90.48	94.21	94.31	97.14
SS-SMALL-16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SS-SMALL-48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VC-C1	0.11	0.00	0.05	0.02	0.03	0.02	0.03	0.00	0.05	0.02	0.01	0.01
VC-EVC	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.05	0.04	0.06	0.04	0.06
VC-FEST	0.50	0.69	0.63	0.68	0.68	0.71	0.21	0.24	0.23	0.24	0.22	0.24
VC-FS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VC-GMM	0.39	0.46	0.45	0.47	0.47	0.48	0.89	0.95	0.92	0.93	0.89	0.92
VC-KPLS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VC-LSP	0.02	0.01	0.01	0.01	0.02	0.02	0.17	0.18	0.17	0.16	0.15	0.17
VC-TVC	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.03	0.03	0.03	0.03

6.58, 0.53, -0.07 and 0.97, respectively. The results for this are presented in the last column of Table VI. The fused countermeasure detects most spoofing attacks, achieving FARs under 1% against all but one spoofing method; it fails to detect SS-MARY. Although the PP countermeasure can discriminate extremely well between human and SS-MARY speech, this ability is not picked up by the fused countermeasure because PP has a low weight. This is because the weights were learned on the development set, which of course only contains known attacks (the first group of 5 countermeasures in Table VI), but the PP countermeasure performs poorly on many of those known attacks, especially the voice conversion ones. Hence, it is given a low weight, and essentially ignored in the fused countermeasure.

D. Spoofing ASV Systems that Employ a Countermeasure

We conducted experiments to evaluate the overall performance of speaker verification systems that include a countermeasure. We only consider the proposed fused countermeasure here, because it exhibited better overall performance than any individual countermeasure. We integrated the fused countermeasure with each of the ASV systems as a post-processing module – as illustrated in Fig. 3 – to reflect the practical use case in which a separately-developed standalone countermeasure is added to an already-deployed ASV system [16] without significant modification of that system.

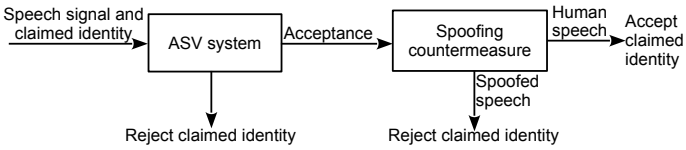

Fig. 3. A speaker verification system with an integrated countermeasure. The integrated system only accepts a claimed identity if it is accepted by the speaker verification system *and* classified as human speech by the countermeasure [16].

A good countermeasure should reduce FARs by rejecting non-human speech. The FAR results of systems with an integrated countermeasure are presented in Table VII. Comparing against the FARs of the ASV systems without a countermeasure in Table V, we can make the following observations. First, the FARs of all ASV systems are reduced dramatically for both male and female speech, and go down from about 70%-100% to below 1% in the face of most types of spoofing attack. This indicates that the fused countermeasure can be effectively integrated with any ASV system without needing additional joint optimisation. Second, the integrated system is robust against attacks from various state-of-the-art statistical parametric speech synthesis and voice conversion systems. However, it is still vulnerable to the unit selection synthesis (SS-MARY) spoofing attack. This suggests that new countermeasures are needed specifically for waveform selection-based spoofing attacks. Third, although our stand-alone ASV systems achieve better performance for male than for female speakers, the integrated systems work equally well for both. In contrast, others have reported integrated systems working better for male speakers than for female speakers [40].

In general, by using the proposed fused countermeasure, the FARs of ASV systems under spoofing attack are reduced significantly. This indicates that the countermeasure is effective in detecting spoofing attacks.

E. Human Versus Machine

To complement the comparisons already presented, we now benchmark automatic (machine-based) methods against speaker verification by human listeners. To do this, three listening tests were conducted: two speaker verification tasks and one spoofing detection task. The first verification task contained only human speech signals, the second verification task contained human speech but all test signals were artificial (synthetic and voice-converted speech). The third task, a detection task, contained both human and artificial speech signals and the goal for the listener was to correctly discriminate these signals. All three tasks covered the 46 target speakers in the evaluation set of the **SAS** corpus.

In order to encourage listeners to engage with the tasks to the best of their ability, they were presented as role-play scenarios. The human listening tasks were designed to be

as similar to the ASV tasks as possible (to facilitate direct comparisons), whilst taking into account listener constraints such as fatigue or memory limitations. Listening protocols were inspired by the ones used in [75] and the experiments were carried out via a web browser. In total, 100 native English listeners took part in the experiments. They were seated in sound-isolated booths and listened to all samples using Beyerdynamic DT 770 PRO headphones. Each listener performed three tasks and, on average, it took about an hour to complete the experiment. We only report results from listeners who completed all sessions in each task.

Task 1: Speaker Verification of Human Speech: In the speaker verification task, listeners were asked to imagine they were responsible for giving people access to their bank accounts. They were informed that they would only have a short recording of a person's voice to base their judgement on. It was stressed that it was important to not give access to "impostors" but equally important that access was given to the "bank account holder".

The listeners were given five sentences from each target speaker to familiarise themselves with the voice. After listening to the training samples, they were given 21 trials to judge as "same" or "different." The trials were pairs of samples which include a reference and a test sample. This was repeated for three different target speakers. In this task, each target speaker was judged by 5 listeners. The number of targets versus non-targets varied per speaker to keep listeners from keeping count for individual speakers. On average there were 10 targets and 11 non-targets per speaker. Genders were not mixed within a trial.

Listeners recognised impostors as genuine targets 2.39% of the time (FAR) while 9.38% of genuine trials were misclassified as impostors (FRR). Comparing with the baseline ASV performance in Table V, the results demonstrate that the speaker verification performance of humans is not as good as that of the best automatic systems. For example, PLDA-5 gives a FAR around 1.5% for both male and female speakers. This finding is similar to that in [76] for the NIST SRE 2008 dataset.

Task 2: Speaker Verification of Artificial Speech: In the second task, listeners were asked to decide whether an artificial voice[17] sounded like the original speaker's voice. The listeners were informed that the artificial voice would sometimes sound quite degraded but were asked to ignore the degradations as much as possible. Additionally, they were told that there would be artificial voices that were supposed to sound like the intended speaker as well as artificial voices that were not supposed to match the original speaker. The task was framed as "Your challenge is to decide which of the artificial voices are based on the 'bank account holder's voice' and which are based on an 'impostor's voice.' "

As in the first task, listeners were given five natural speech samples from the intended speaker to familiarise themselves with the voice. After listening to these training samples, subjects were presented with pairs of reference and test samples to judge as "same" or "different." It was made clear to the listeners that the test sample would be an artificial voice. Each target speaker was judged by 5 listeners. For each target speaker there were 65 trials (13 systems, each presented 5 times). On average there were 39 targets and 26 non-targets per speaker. Once again gender was not mixed within any of the trials.

The results are presented in Table VIII (second column). The acceptance rate is not directly comparable with the automatic results presented in Table V but the relative differences across spoofing algorithms are comparable[18].

It can be observed that SS-MARY gives the highest acceptance rate (i.e., listeners said that it sounded like the original speaker), while VC-C1 gives the lowest acceptance rate – this pattern is similar to that in the ASV results where SS-MARY achieves relatively high FARs and VC-C1 relatively low FARs. The results also indicate that spoofing systems that use more training data generally achieve higher acceptance rates with human listeners, mirroring what we saw earlier in the ASV results in Section V-B. An interesting difference between the ASV and human listener results is that, for human listeners, the use of higher sampling rate speech by some spoofing systems (SS-SMALL-48, SS-LARGE-48) leads to a lower acceptance rate than for lower sampling rate training data (SS-SMALL-16, SS-LARGE-16). This suggests that, whilst these types of spoofing systems (SS: statistical parametric speech synthesis) are able to generate information above 8 kHz that contributes to improved naturalness [51], listeners judge it as being more dissimilar to the natural speaker. This similarity observation is different from that in [51], where speaker-dependent speech synthesis is examined. An informal listening test gives the impression that SS-LARGE-48/SS-SMALL-48 produces more natural speech than SS-LARGE-16/SS-SMALL-16, as expected. However, as the reference target speech is a clean recording without any distortion, we speculate that it is more challenging for listeners to decide on the speaker similarity of the poor quality, buzzy-sounding speech of SS-LARGE-16/SS-SMALL-16.

Task 3: Detection: In the final task, listeners were asked to judge whether a speech sample was a recording of a human voice, or a sample of an artificial voice. The challenge to the listeners was formulated as: "Imagine an impostor trying to gain access to a bank account by mimicking a person's voice using speech technology. You must not let this happen. Your challenge in this final section is to correctly tell whether or not the sample is of a human or of a machine."

Listeners were again given some training speech signals. They listened to five samples of human speech from one speaker (not present in the detection task) and five examples of artificial speech generated using five known spoofing systems.

[17]Artificial was explained to the listeners as being "produced by a machine, computer-generated, for example a synthetic voice".

[18]In Task 2, the acceptance rate means the percentage of genuine speech recognised as the original speaker. The genuine speech is artificial speech using the target speaker's voice as the reference for adaptation or voice conversion, and the impostor speech is also artificial speech but using a non-target speakers voice as the reference for adaptation or voice conversion. When computing the acceptance rate, zero is used as the threshold. On the other hand, the FAR in Table IV is the percentage of spoofed trials accepted as genuine. When computing the FAR, the threshold is determined at the EER point on the non-spoofed trials.

TABLE VIII
TASK 2 –SPEAKER VERIFICATION (ARTIFICIAL)– AND TASK 3 –SPOOFING DETECTION– RESULTS.

	Task 2: Speaker Verification Acceptance rate	Task 3: Spoofing Detection Detection Error Rate
Human	-	11.94
SS-SMALL-16	35.33	5.48
SS-SMALL-48	32.19	7.86
SS-LARGE-16	39.46	5.71
SS-LARGE-48	36.18	8.10
SS-MARY	76.07	8.10
VC-GMM	30.63	13.57
VC-KPLS	29.06	6.90
VC-TVC	20.51	5.48
VC-EVC	21.23	8.10
VC-FS	35.47	4.29
VC-C1	7.26	23.81
VC-FEST	28.63	6.90
VC-LSP	23.36	7.38

At this point, the listeners were informed that the training samples did not cover all possible types of artificial speech. In Task 3, there were 130 samples (65 human, 65 artificial (13×5)), and those samples were randomly selected from the evaluation set for each listener. 84 listeners participated in the test.

The human detection results are presented in Table VIII (third column). In general, human listeners detect spoofing less well than most of the automatic approaches presented in Table VI. For most spoofing systems, the automatic approaches give FARs below 1%, while human listeners have FARs above 4%. However, humans are much better than any of the automatic countermeasures (except PP) in detecting SS-MARY. Most of the countermeasures exhibit FARs in excess of 80% for SS-MARY, while the FAR of human listeners is only 8%.

VI. DISCUSSION AND FUTURE WORK

In this section, we summarise the findings in this work, and also discuss some of its limitations. Both the findings and the limitations suggest areas needing further research.

A. Research Findings

The main findings from this study are:

- All three classical ASV systems: GMM-UBM, JFA and PLDA systems are vulnerable to all the spoofing methods considered, with the exception of VC-C1. This confirms the findings of previous studies that only considered one or two spoofing algorithms. This also shows the importance of developing spoofing countermeasures to secure ASV systems.
- The effectiveness of speech synthesis and voice conversion spoofing are comparable. Previous studies employed various databases for each attack which made direct comparisons of effectiveness across attacks difficult or impossible. The standardised protocol that we propose here, using our **SAS** database, allows direct comparisons.
- When higher sampling rate and/or more training data are available to train spoofing systems, FARs of ASV systems increase significantly, as expected. This indicates that ASV systems are more vulnerable to attackers who have access to better quality and/or greater quantity of training data.
- Generally, the spoofing countermeasures proposed in this work perform well in detecting statistical parametric speech and voice conversion attacks. However, they mostly fail to detect rather straightforward waveform concatenation, as in the case of the SS-MARY attack. Because SS-MARY directly concatenates waveforms in the time-domain, the resulting spoofed speech has no distortions in the phase domain (except perhaps at the concatenation points); so, phase-based countermeasures are not a good way to detect such a spoofing attack.
- ASV systems have reached a point where they routinely outperform ordinary humans[19] on speaker recognition and spoofing detection tasks. However, humans are still better able to detect waveform concatenation. An obvious practical approach at the current time, for example in call-centre applications, would be to combine the decisions of both human and automatic systems.

B. Limitations and Future Directions

We suggest future work in ASV spoofing and countermeasures along the following lines:

- **More diverse spoofing materials:** The current **SAS** database is biased towards the STRAIGHT vocoder, and only one type of unit selection system was used to generate the waveform concatenation materials. Moreover, replay attack – which does not require any speech processing knowledge on the part of the attacker – was not considered here. A generalised countermeasure should be robust against all spoofing algorithms and any vocoder. The development of generalised countermeasures might be accelerated by collecting more diverse spoofing materials. As the amount of spoofing materials increases, ASV systems can access more representative prior information about spoofing, and the security of ASV systems should be enhanced as a result.
- **Truly generalised countermeasures:** The proposed countermeasures did not generalise well to unknown attacks, and in particular to the SS-MARY attack. This is because the proposed countermeasures were biased towards detecting phase artefacts. To detect the SS-MARY attack or similar waveform concatenation attacks, we suggest further development of pitch pattern-based countermeasures. Discontinuity detection for concatenative speech synthesis [77] might also be useful in inspiring novel countermeasures against such attacks. Lastly, novel system fusion methods might also be a way to implement generalised countermeasures. A good fusion method should be able to benefit from all the individual countermeasures. Our proposed fusion method failed to take advantage of the strengths of the pitch pattern countermeasure, for example.

[19]It would be interesting in the future to use either 'super recognisers' or forensic speech scientists, if we could access sufficient numbers of such listeners.

- **Noise or channel robustness:** The work here delberately focussed on clean speech without significant noise or channel effects. To make the proposed countermeasures appropriate for practical applications, it would of course be important to take channel and noise issues into consideration.
- **Text-dependent ASV:** The current work assumes text-independent speaker verification. To make systems suitable for other voice authentication applications, spoofing countermeasures for text-dependent ASV must also be developed.

VII. Conclusions

All existing literature that we are aware of in the areas of ASV spoofing and anti-spoofing, report results for just one or two spoofing algorithms, and generally assumes prior knowledge of the spoofing algorithm(s) in order to implement matching countermeasures. As discussed in [8], the lack of a large-scale, standardised dataset and protocol was a fundamental barrier to progress in this area. We hope that this situation is now rectified, by our release of the standard dataset **SAS**, combined with the benchmark results presented in this paper.

To acheive this, speech synthesis, voice conversion, and speaker verification researchers worked together to develop state-of-the-art systems from which to generate spoofing materials, and thus to develop countermeasures. The **SAS** corpus developed in this work is publicly released under a CC-BY license [78]. We hope that the availability of the **SAS** corpus will facilitate reproducible research and as a consequence drive forward the development of novel generalised countermeasures against speaker verification system spoofing attacks.

VIII. Acknowledgements

The authors would like to thank: Dr. Tomi Kinnunen and Dr. Nicolas Evans for their valuable comments which have improved the speaker verification and countermeasure protocols; Dr. Ling-Hui Chen and Ms. Li-Juan Liu for their assistance in generating spoofing materials; the three anonymous reviewers for their valuable comments and suggestions to improve the manuscript.

References

[1] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, and S. King, "SAS: A speaker verification spoofing database containing diverse attacks," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[2] M. Wester, Z. Wu, and J. Yamagishi, "Human vs machine spoofing detection on wideband and narrowband data," in *Proc. Interspeech*, 2015.

[3] P. Golden, "Voice biometrics–the Asia Pacific experience," *Biometric Technology Today*, vol. 2012, no. 4, pp. 10–11, 2012.

[4] M. Khitrov, "Talking passwords: voice biometrics for data access and security," *Biometric Technology Today*, vol. 2013, no. 2, pp. 9–11, 2013.

[5] B. Beranek, "Voice biometrics: success stories, success factors and what's next," *Biometric Technology Today*, vol. 2013, no. 7, pp. 9–11, 2013.

[6] 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*, February 2013.

[7] W. Meng, D. Wong, S. Furnell, and J. Zhou, "Surveying the development of biometric user authentication on mobile phones," *IEEE Communications Surveys and Tutorials*, 2015.

[8] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[9] 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.

[10] 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.

[11] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13–31, 2015.

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

[13] Z. Wu, S. Gao, E. S. Chng, and H. 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.

[14] Z. Wu and H. Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," *Multimedia Tools and Applications*, 2015.

[15] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2010.

[16] 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.

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

[18] 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.

[19] 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.

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

[21] Z. Wu and H. Li, "Voice conversion versus speaker verification: an overview," *APSIPA Transactions on Signal and Information Processing*, vol. 3, p. e17, 2014.

[22] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 1999.

[23] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. Interspeech*, 2000.

[24] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proc. the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[25] B. L. Pellom and J. H. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

[26] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, "Voice forgery using ALISP: indexation in a client memory," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[27] D. Matrouf, J.-F. Bonastre, and C. Fredouille, "Effect of speech transformation on impostor acceptance," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[28] 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.

[29] P. L. De Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Proc. Interspeech*, 2012.

[30] R. D. McClanahan, B. Stewart, and P. L. De Leon, "Performance of i-vector speaker verification and the detection of synthetic speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[31] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[32] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[33] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *Proc. Interspeech*, 2013.

[34] F. Alegre, A. Amehraye, and N. 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.

[35] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. Interspeech 2012*, 2012.

[36] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "A cross-vocoder study of speaker independent synthetic speech detection using phase information," in *Proc. Interspeech*, 2014.

[37] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Trans. on Information Forensics and Security*, vol. 10, no. 4, pp. 810–820, 2015.

[38] 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.

[39] E. Khoury, T. Kinnunen, A. Sizov, Z. Wu, and S. Marcel, "Introducing i-vectors for joint anti-spoofing and speaker verification," in *Proc. Interspeech*, 2014.

[40] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispoofing in the i-vector space," *IEEE Trans. on Information Forensics and Security*, vol. 10, no. 4, pp. 821–832, 2015.

[41] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2014. [Online]. Available: http://www.zhizheng.org/papers/asvSpoof_eval_plan.pdf

[42] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech*, 2015.

[43] C. S. Greenberg, A. F. Martin, L. Brandschain, J. P. Campbell, C. Cieri, G. R. Doddington, and J. J. Godfrey, "Human assisted speaker recognition in nist sre10." in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2010.

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

[45] 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 Communication*, vol. 27, pp. 187–207, 1999.

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

[47] S. Watanabe, A. Nakamura, and B.-H. Juang, "Structural bayesian linear regression for hidden Markov models," *Journal of Signal Processing Systems*, vol. 74, no. 3, pp. 341–358, 2014.

[48] 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.

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

[50] 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.

[51] J. Yamagishi and S. King, "Simple methods for improving speaker-similarity of HMM-based speech synthesis," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[52] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[53] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation." in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 1994.

[54] 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.

[55] 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.

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

[57] 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.

[58] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based unit selection for voice conversion utilizing temporal information," in *Proc. Interspeech*, 2013.

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

[60] E. Helander, H. Silén, T. Virtanen, and M. 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.

[61] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

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

[63] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. Prince, "Probabilistic models for inference about identity," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 144–157, January 2012.

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

[65] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1. 0: A MATLAB toolbox for speaker recognition research," in *IEEE Signal Processing Society Speech and language Technical Committee Newsletter*, November 2013.

[66] P. Rajan, A. Afanasyev, V. Hautamäki, and T. Kinnunen, "From single to multiple enrollment i-vectors: Practical plda scoring variants for speaker verification," *Digital Signal Processing*, vol. 31, pp. 93–101, 2014.

[67] K. K. Paliwal and L. D. Alsteris, "Usefulness of phase spectrum in human speech perception." in *Proc. Interspeech*, 2003.

[68] R. M. Hegde, H. Murthy, V. R. R. Gadde *et al.*, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.

[69] P. Rajan, S. H. K. Parthasarathi, and H. A. Murthy, "Robustness of phase based features for speaker recognition," in *Proc. Interspeech*, 2009.

[70] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[71] P. Taylor, *Text-to-speech synthesis*. Cambridge University Press, 2009.

[72] A. Ogihara, H. Unno, and A. Shiozakai, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 88, no. 1, pp. 280–286, jan 2005.

[73] N. Brummer, L. Burget, J. H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. Van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[74] V. Hautamaki, T. Kinnunen, F. Sedlák, K. A. Lee, B. Ma, and H. Li, "Sparse classifier fusion for speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 21, no. 8, pp. 1622–1631, 2013.

[75] A. Schmidt-Nielsen and T. H. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data," *Digital Signal Processing*, vol. 10, no. 1, pp. 249–266, 2000.

[76] V. Hautamäki, T. Kinnunen, M. Nosratighods, K. A. Lee, B. Ma, and H. Li, "Approaching human listener accuracy with modern speaker verification," in *Proc. Interspeech*, 2010.

[77] Y. Pantazis, Y. Stylianou, and E. Klabbers, "Discontinuity detection in concatenated speech synthesis based on nonlinear speech analysis," in *Proc. Interspeech*, 2005.

[78] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, Z.-H. Ling, and S. King, "Spoofing and Anti-Spoofing (SAS) corpus v1.0," 2015. [Online]. Available: http://dx.doi.org/10.7488/ds/252

Zhizheng Wu received his Ph.D. from Nanyang Technological University, Singapore. He was a visiting intern at Microsoft Research Asia as a visiting intern and a visiting researcher at the University of Eastern Finland. Since 2014, he is a research fellow in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh. He received the best paper award at the Asia Pacific Signal and Information Processing Association Annual Submit and Conference (APSIPA ASC) 2012, co-organised the first Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) at Interspeech 2015, delivered a tutorial on "Spoofing and Anti-Spoofing: A Shared View of Speaker Verification, Speech Synthesis and Voice Conversion" at APSIPA ASC 2015 and co-organised the first Voice Conversion Challenge (VCC 2016).

Phillip De Leon (SM '03) received the B.S. Electrical Engineering and the B.A. in Mathematics from the University of Texas at Austin, in 1989 and 1990 respectively and the M.S. and Ph.D. degrees in Electrical Engineering from the University of Colorado at Boulder, in 1992 and 1995 respectively. In 2002, he was a visiting professor in the Department of Computer Science at University College Cork, Ireland. In 2008, he was selected by the U. S. State Department as a Fulbright Faculty Scholar and served as a visiting professor at Technical University in Vienna (TU-Wien). He currently holds the Paul W. and Valerie Klipsch Distinguished Professorship in Electrical and Computer Engineering at NMSU and directs the Advanced Speech and Audio Processing Laboratory. He has co-authored over 70 refereed papers in international journals and conferences. His research interests include machine learning, speaker recognition, speech enhancement, and time-frequency analysis. He is a member of the Industrial Digital Signal Processing Technical Committee (IDSP-TC).

Cenk Demiroglu obtained the B.S. degree in electrical and electronics engineering from Bogazici University in 1999, the M.S. degree from the Electrical and Electronics Engineering Department of University of Nebraska, Lincoln, in 2001, and the Ph.D. degree from the Electrical and Computer Engineering Department of the Georgia Institute of Technology in 2005. After the Ph.D. degree, he was with the R&D groups of speech technology companies in the USA for five years. He played lead roles in the development of large-vocabulary speech recognition systems for three years and the development of embedded text-to-speech synthesis systems for two years. He joined the Ozyegin University Electrical and Electronics Engineering Department in 2009 as an Assistant Professor. His research and consulting activities are focused on speech synthesis, speech recognition, and speaker verification.

Ali Khodabakhsh obtained the B.Sc. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2011, and the M.Sc. degree in computer science from Ozyegin University, Istanbul, Turkey, in 2015. His research interests include speaker recognition, spoofing and anti-spoofing, and deep learning.

Simon King (M'95–SM08–F15) holds M.A.(Cantab) and M.Phil. degrees from Cambridge and a Ph.D. from Edinburgh. He has been with the Centre for Speech Technology Research at the University of Edinburgh since 1993, where he is now Professor of Speech Processing and the director of the centre. His interests include speech synthesis, recognition and signal processing and he has around 200 publications across these areas. He has served on the ISCA SynSIG board and currently co-organises the Blizzard Challenge. He has previously served on the IEEE SLTC and as an associate editor of IEEE Transactions on Audio, Speech and Language Processing, and is currently an associate editor of Computer Speech and Language.

Zhen-Hua Ling (M10) received the B.E. degree in electronic information engineering, M.S. and Ph.D. degree in signal and information processing from University of Science and Technology of China, Hefei, China, in 2002, 2005, and 2008, respectively. From October 2007 to March 2008, he was a Marie Curie Fellow at the Centre for Speech Technology Research (CSTR), University of Edinburgh, UK. From July 2008 to February 2011, he was a joint Postdoctoral Researcher at the University of Science and Technology of China and iFLYTEK Co., Ltd., China. He is currently an Associate Professor at the University of Science and Technology of China. He also worked at the University of Washington, USA, as a Visiting Scholar from August 2012 to August 2013. His research interests include speech processing, speech synthesis, voice conversion, speech analysis, and speech coding. He was awarded IEEE Signal Processing Society Young Author Best Paper Award in 2010.

Daisuke Saito received the B.E., M.S., and Dr. Eng. degrees from the University of Tokyo, Tokyo, Japan, in 2006, 2008, and 2011, respectively. From 2010 to 2011, he was a Research Fellow (DC2) of the Japan Society for the Promotion of Science. He is currently an Assistant Professor in the Graduate School of Information Science and Technology, University of Tokyo. He is interested in various areas of speech engineering, including voice conversion, speech synthesis, acoustic analysis, speaker recognition, and speech recognition. Dr. Saito is a member of the International Speech Communication Association (ISCA), the Acoustical Society of Japan (ASJ), the Information Processing Society of Japan (IPSJ), the Institute of Electronics, Information and Communication Engineers (IEICE), and the Institute of Image Information and Television Engineers (ITE). He received the ISCA Award for the best student paper of INTERSPEECH 2011, the Awaya Award from the ASJ in 2012, and the Itakura Award from ASJ in 2014.

Bryan Stewart received the B.S., M.S. and Ph.D. degrees in Electrical Engineering from New Mexico State University in 2004, 2006, and 2016 respectively. In 2006, he started working in the System Engineering Directorate at White Sand Missile Range, NM on Unmanned Autonomous Systems Test and Evaluation. In 2010, he started working for the Naval Surface Warfare Center Port Hueneme Division and leads a team of engineers in developing augmented reality, prognostics, secure wireless, and cybersecurity capability for the surface Navy. His interests include machine learning, natural language processing, and cybersecurity.

Tomoki Toda received his B.E. degree from Nagoya University, Japan, in 1999 and his M.E. and D.E. degrees from Nara Institute of Science and Technology (NAIST), Japan, in 2001 and 2003, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science from 2003 to 2005. He was then an Assistant Professor (2005-2011) and an Associate Professor (2011-2015) at NAIST. From 2015, he has been a Professor in the Information Technology Center at Nagoya University. His research interests include statistical approaches to speech processing. He received more than 10 paper/achievement awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal).

Mirjam Wester received the Ph.D. degree from the University of Nijmegen, the Netherlands, in 2002. She was a visiting researcher at ICSI, Berkeley, CA from March 2000 to March 2001. Since 2003 she has been a Research Fellow at the Centre for Speech Technology Research, University of Edinburgh, UK. Her research interests focus on taking knowledge of human speech production and perception and applying it to speech technology.

Junichi Yamagishi (Senior Member, IEEE) was awarded a Ph.D. by the Tokyo Institute of Technology in 2006 for a thesis that pioneered speaker-adaptive speech synthesis and was awarded the Tejima Prize as the best Ph.D. thesis of Tokyo Institute of Technology in 2007. He is an Associate Professor at the National Institute of Informatics in Japan. He is also a Senior Research Fellow in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK. Since 2006, he has authored and co-authored about 150 refereed papers in international journals and conferences. He was awarded the Itakura Prize from the Acoustic Society of Japan, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan, and the Young Scientists Prize from the Minister of Education, Science and Technology in 2010, 2013, and 2014, respectively. He was a scientific committee and area coordinator for Interspeech 2012. He was one of organizers for the above special sessions on ”Spoofing and Countermeasures for Automatic Speaker Verification at Interspeech 2013 and ASVspoof at Interspeech 2015. He has been a member of the Speech & Language Technical Committee (SLTC) and an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing.