Front-End for Anti-Spoofing Countermeasures in Speaker Verification: Scattering Spectral Decomposition

Article *in* IEEE Journal of Selected Topics in Signal Processing · December 2016

DOI: 10.1109/JSTSP.2016.2647202

CITATIONS
8

READS
79

4 authors, including:

Kaavya Sriskandaraja
UNSW Sydney
12 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Vidhyasaharan Sethu
UNSW Sydney
80 PUBLICATIONS **580** CITATIONS
SEE PROFILE

Haizhou Li
National University of Singapore
641 PUBLICATIONS **7,529** CITATIONS
SEE PROFILE

Front-End for Antispoofing Countermeasures in Speaker Verification: Scattering Spectral Decomposition

Kaavya Sriskandaraja, Vidhyasaharan Sethu, *Member, IEEE*, Eliathamby Ambikairajah, *Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

***Abstract*—As speaker verification is widely used as a means of verifying personal identity in commercial applications, the study of antispoofing countermeasures has become increasingly important. By choosing appropriate spectral and prosodic feature mapping, spoofing methods based on voice conversion and speech synthesis are both capable of deceiving speaker verification systems that typically rely on these features. Consequently alternative front-ends are required for effective spoofing detection. This paper investigates the use of the recently proposed hierarchical scattering decomposition technique, which can be viewed as a generalization of all constant-Q spectral decompositions, to implement front-ends for stand-alone spoofing detection. The coefficients obtained using this decomposition are converted to a feature vector of Scattering Cepstral Coefficients (SCCs). We evaluate the performance of SCCs on the recent spoofing and Antispoofing (SAS) corpus as well as the ASVspoof 2015 challenge corpus and show that SCCs are superior to all other front-ends that have previously been benchmarked on the ASVspoof corpus.**

***Index Terms*—Anti-spoofing, automatic speaker verification, modulation spectrum, scattering spectrum, spoofing countermeasures, spoofing detection, wavelet decomposition.**

I. Introduction

SPEAKER verification, also known as voice authentication, aims to verify a claimed identity based on the person's speech. Compared to other forms of biometric identity verification such as face or finger print based methods, voice is an appealing modality since speech is the primary mode of communication and it allows for remote authentication over common communication channels such as telephone lines and voice over IP. Speaker verification systems are typically automatic systems that accept or reject a claimed identity based on a speech utterance from the claimant and can be broadly categorised into text-dependent or text-independent systems. Text-dependent systems use fixed phrases for verification whereas text-independent systems operate on arbitrary utterances. While text-dependent systems currently offer high verification accuracy, text-independent systems don't rely on fixed passphrases thereby making possible speaker verification over spontaneous speech.

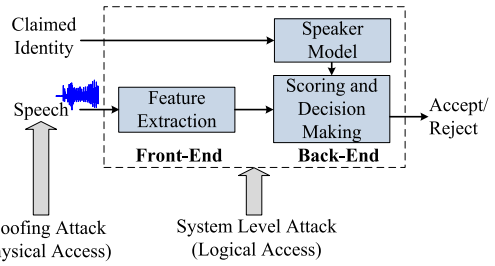

Fig. 1. Schematic diagram of attacks on speaker verification systems.

Speaker verification is now available in smartphone logical access scenarios [1] in e-commerce [2] and in mobile banking [3]. Given the remote nature of most speaker verification systems (without face to face contact) protecting against spoofing attacks, which aim to impersonate valid users, is of fundamental concern. The popularity of social media has also made it relatively easy for someone with malicious intent to steal voice samples from intended target speakers [4], [5]. Spoofing attacks pose a serious threat, in particular to the banking industry, where mass attacks could cause financial losses and more importantly a loss of public confidence in financial systems. This paper focuses on anti-spoofing methods/countermeasures for text-independent speaker verification systems.

Speaker verification systems (Fig. 1) typically operate by extracting speaker characteristics from a speech utterance and comparing them to the speaker model of the claimed identity. Attacks can be targeted either at the speech input (physical access) level or at the system (logical access) level. As the latter attacks require system level access, they are less of a threat and are not considered in this paper.

The vulnerability of state-of-the-art speaker verification systems was reported in recent studies [5], [6]. Spoofing methods based on physical access fall into one of four categories,

Manuscript received August 15, 2016; revised November 15, 2016; accepted December 16, 2016. Date of publication December 30, 2016; date of current version May 11, 2017. The guest editor coordinating the review of this paper and approving it for publication was Prof. Phillip L. DeLeon.

K. Sriskandaraja and V. Sethu are with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, N.S.W. 2052, Australia (e-mail: k.sriskandaraja@unsw.edu.au; v.sethu@unsw.edu.au).

E. Ambikairajah is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, N.S.W. 2052, Australia, and also with the Data61, ATP, Eveleigh, N.S.W. 2015, Australia (e-mail: ambi@ee.unsw.edu.au).

H. Li is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, and also with the Institute for Infocomm Research, A*STAR, Singapore 138632 (e-mail: eleliha@nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2016.2647202

namely, impersonation [7], replay [8], speech synthesis [9] and voice conversion [10], [11] Here, impersonation refers to one person mimicking another; replay refers to recording the speech of a person and playing it back to a speaker verification system; speech synthesis refers to text-to-speech waveform generation; and voice conversion refers to an automatic system that takes in speech of one person and re-synthesizes it to sound like that of another.

Among the four categories, impersonation is unlikely to be effective [12] and is not practical for large scale attacks. Furthermore, while replay attacks are somewhat straightforward to carry out, they are primarily a threat only to text-dependent systems. Voice conversion and speech synthesis attacks are however serious threats due to widely available state-of-the-art open source software making them both effective and accessible.

Almost all speaker verification systems extract cepstral features, typically mel frequency cepstral coefficients (MFCCs), and model the distributions of these cepstral features to represent speakers. Consequently, effective spoofing methods based on voice conversion and speech synthesis are designed to produce speech that leads to similar distributions of cepstral coefficients as the target speaker. It is therefore logical to use alternative features that capture a richer representation of the spectral content of speech in order to better distinguish between genuine and spoofed speech. Furthermore, while improvements can also be made to the back-end modelling and classification techniques, these rely on the information captured at the feature level and therefore feature domain analyses should precede other back-end developments.

In Section II, an overview of state-of-the-art speaker verification systems is provided; Section III discusses current countermeasures to spoofing attacks; Section IV introduces multi-level spectral features based on the scattering spectrum for these countermeasures; and Section V reports experimental results.

II. Speaker Verification Systems

State-of-the-art speaker verification systems generally comprise of a number of components (Fig. 2) as outlined below.

Front-End Processing: The front-end is designed to extract discriminative features. Among these, MFCCs [13] are the most widely used features along with their velocity and acceleration features, also known as Δ s and $\Delta - \Delta$ s respectively, which are computed from consecutive frames of the original MFCCs [14]. Recently bottleneck features based on deep neural network (DNN) architectures are gaining popularity in addition to MFCCs [15]. Typically silences and pauses between words are removed as part of the front-end using a voice activity detector [16].

Modelling and Compensation: The features extracted from an input test utterance are compared to a speaker model corresponding to the claimed speaker identity and a background model representing all speakers other than the claimed speaker. Traditionally the GMM-UBM (Gaussian Mixture Model – Universal Background Model) approach has been used for speaker modelling. However, an alternative that is rapidly gaining interest is the use of deep neural networks (DNNs) [17]. GMMs can be parametrised as supervectors which in turn can be mapped into a much lower dimensional vectors referred to as 'i-vectors' [18]. If DNNs are used to model speech acoustics in the speaker modelling stage, the DNN output posteriors can be mapped to i-vectors [19]. Finally, the classification/scoring stage of a speaker verification system compares an i-vector estimated from the test speech features to both the claimed speaker model and the background model and generates a score. Currently the most widely used approach to this scoring is the use of a Gaussian probabilistic linear discriminant analysis (GPLDA) [20].

Normalisation: The 'channel' for the speech signal plays a critical role in speaker verification systems. Here, the term channel refers to background noise as well as the spectral characteristics of the recording setup including microphone and telephone channel characteristics. Speaker verification performance can significantly deteriorate if there is a mismatch of channel conditions between training and test data. Channel normalisation is typically carried out at three different levels in speaker verification systems to reduce this mismatch – at the feature level, the speaker model level and the score level (Fig. 2).

III. Spoofing and Countermeasures

A. Spoofing

The vulnerability of speaker verification systems to spoofing attacks has been shown in recent work [7], [8], [10]–[12]. In this paper, we study spoofing countermeasures against speech synthesis and voice conversion.

Speech Synthesis: Also known as text-to-speech (TTS), speech synthesis refers to systems that generate speech corresponding to any input text [21]–[24]. Among these, the methods that lead to high quality natural sounding speech generally take one of two approaches, either 'unit selection' methods [25], [26] or 'statistical parametric' methods [24], [27].

Voice Conversion: The aim of voice conversion is to transform speech from one speaker to another. Generally voice conversion is carried out in 3 stages. The input speech is firstly analysed into features, which are then transformed to match the characteristics of the target speaker, and finally the transformed features are re-synthesized into speech with a vocoder [28], [29]. The transformation of the features normally involves spectral and prosodic mapping that is learned from both source and target speakers [30]. In the context of spoofing, all state-of-the-art speaker verification systems operate on either spectral and/or prosodic feature spaces. Therefore, by choosing appropriate spectral and prosodic mapping in voice conversion, converted speech can effectively deceive a speaker verification system even with audible artefacts that human ears can detect [12]. Such transformation methods include vector quantisation [31], joint density GMMs [32], tensor based methods [33], spectral warping [34], [35], etc.

B. Countermeasures

Typically we compensate non-speaker variability in a speaker verification system to improve its robustness. As a result, more robust speaker verification may become more vulnerable to

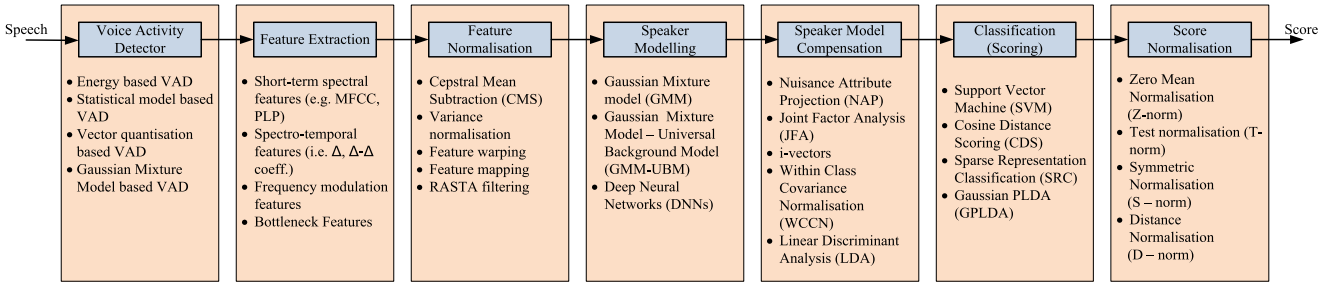

Fig. 2. Summary of components of speaker verification systems.

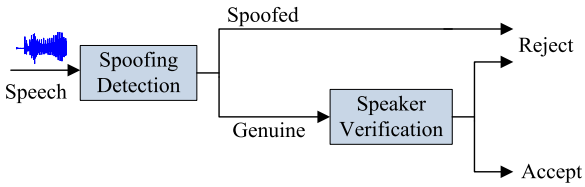

Fig. 3. Stand-alone countermeasure for spoofing detection.

spoofing [12]. In view of the fact that speaker verification and anti-spoofing systems may have conflicting optimisation criteria, they are usually designed separately in two independent systems.

There are two broad approaches to incorporating anti-spoofing countermeasures [12]. One approach is to fuse the decision of speaker verification and spoofing detection scores, that we call the parallel approach. The alternate approach is to have a standalone countermeasure that pipelines with the speaker verification system in series (Fig. 3), Both approaches have their merits: the parallel system allows for a shared front-end [36] which could be computationally more efficient; while the standalone countermeasure can operate independently without modifying the speaker verification systems [12].

Given that most voice conversion and speech synthesis methods aim to match the characteristics of the target speaker that are captured by the standard mel cepstral front-ends employed in speaker verification, we consider that an alternative front-end that captures complementary information could be more effective for spoofing detection. The different front-ends that have been proposed for spoofing detection can be broadly categorised as either spectral and cepstral features, phase based features, or linear prediction based features.

We consider that voice conversion and speech synthesis techniques focus very much on the spectral envelope rather than the fine details. Therefore, the spectral and cepstral features that capture the fine details may be more useful in detecting spoofing attacks than MFCCs. Furthermore it has been shown that the spectral regions of interest in speaker verification are distinct from those of interest in spoofing detection [37]. Next are some recently studied spectral and cepstral features.

Higher order mel-cepstral coefficients (MCEP) [38] have also been proposed to detect speech synthesised by HMM-based spoofing attacks and mel frequency principal coefficients that make use of Eigenvector bases instead of discrete cosine bases have been shown to outperform traditional MFCCs [39]. Log magnitude spectra have also been proposed as a front-end with the aim of capturing spectral details and harmonic structure in order to detect spoofed speech [39], [40]. Spectral centroid magnitude and frequency features that may be complementary to MFCC features have also been proposed for use in spoofing detectors [41]. Currently, the most promising cepstral features for spoofing detection are those using cochlear filters [42] and constant-Q filters [43], with the constant-Q cepstral coefficients exhibiting the most promising performance [43]. Features based on modulation spectra and other long term spectro-temporal features have also shown to be promising for spoofing detection [44].

In addition to these spectral and cepstral features, other features based on phase and linear prediction have also been investigated for spoofing detection. Some of these include phase based features such as modified group delay features [40], [45], base-band phase difference [40], relative phase shift [41], instantaneous frequency estimates [45], pitch synchronous phase spectra [40] and cosPhasePC features [39], [45], and linear prediction based features [46] such as frequency domain linear prediction features [41] and residual log magnitude spectra [47]. Finally, it has been suggested that the prosody of synthetic speech is generally not the same as that of natural speech [48].

The above mentioned front-ends typically work with suitable back-ends, such as Gaussian mixture models or support vector machines to form spoofing detection systems. There are also recent studies of deep neural networks (DNNs) in the context of spoofing detection. DNNs have been used both in the front-end, in the form of bottleneck features [49] and the so called s-vector or spoofing-vector [49], and the back-end as a classifier [50].

IV. Scattering Spectrum Front End

Among the number of different front-ends that have been proposed for spoofing detection (Section-III.B), the constant-Q cepstral coefficients [43] and the cochlear filters [42], [51] have exhibited the most promising performance. Both of these front-ends employ similar filter-banks and may be considered within the unified framework of the recently proposed scattering spectrum [52]. The scattering spectrum is also closely related to modulation spectra which have also been used in spoofing detection systems [44]. Finally the widely used MFCCs can also be cast within the scattering spectrum framework [53]. Consequently, the scattering spectrum can offer a coherent framework within which front-ends for spoofing detection can be studied and designed.

A. Scattering Decomposition

The scattering transform [52] is a hierarchical spectral decomposition of a signal based on (a) wavelet filter-banks (constant-Q filter-banks), (b) modulus operator (absolute value), and (c) averaging. As shown in Fig. 5, each level of decomposition comprises of running the input signal (outputs from the previous levels) through the wavelet filter-bank and taking the absolute value of filter outputs leading to a scalogram. The scattering coefficients at that level are estimated by windowing the scalogram signals and computing the average value within these windows.

At each level of decomposition, k, a constant-Q filter-bank of N_k filters constitutes a wavelet transform given by a set of wavelets denoted by $\{\psi_{k\,j}[n];\ \ j = 1,\ 2, \ldots, N_k\}$, where $\psi_k[n]$ is the chosen mother wavelet (bandpass filter) and the dilated wavelets are:

$$\psi_{k\,j}\,[n] = \frac{1}{a}\psi_k\left[\frac{nT}{a}\right],\ \ a = 2^{-j/Q_k} \tag{1}$$

where, a is the scaling factor; T is the sampling period; n is the sample index and Q_k denotes the number of filters per octave in the filter-bank at the k^{th} level of decomposition.

The scalogram is computed by taking the absolute values of the outputs of the N_k filters in this constant-Q filter bank and is given by the set of signals $\{p_j[n]\}$, where

$$p_j\,[n] = |(\psi_j * x)\,[n]| \tag{2}$$

Here, $x[n]$ denotes the input signal being decomposed at that level and $|\cdot|$ denotes modulus operator of taking the absolute value.

The scattering coefficients at each level are then obtained by framing $p_j[n]$ using rectangular windows, $\phi[n]$, of length M (corresponding to a low pass filter with a cut-off frequency of π/M) and taking the average value within these frames to obtain the j^{th} scattering coefficient corresponding to the signal $x[n]$ at the k^{th} level of decomposition. i.e.,

$$S_k^{(j)}\,[m] = (\phi * p_j)\,[mM]\,;\ for\ k > 1 \tag{3}$$

where, $S_k^{(j)}[m]$ denotes the scattering coefficient corresponding to the j^{th} bandpass filter, $\psi_j[n]$, at the k^{th} level of decomposition and m denotes the frame index.

Given a speech signal, $s[n]$, and a window size of M, the scattering coefficients for two levels of decomposition (Fig. 5) are computed hierarchically as follows, starting with a single dimensional zeroth level scattering coefficient $S_0[m]$.

$$S_0\,[m] = (\phi * s)\,[mM] \tag{4}$$

The first level scattering coefficients, $S_1[m]$, comprise N_1 coefficients per frame corresponding to an N_1 filter constant-Q filter-bank and is computed as:

$$S_1[m] = \left\{S_1^{(j)}[m]; \forall j\right\} = \left\{\left(\phi * \left|\left(\psi_{1_j} * s\right)\right|\right)[mM];\ \forall j\right\} \tag{5}$$

where, $j = 1,\ 2,\ \ldots N_1$, denotes the j^{th} filter in the filter-bank.

The log scattering coefficients, $\log S_1[m]$, obtained from a speech signal, $s[n]$, along with the corresponding log-scalograms, $\{\log|\psi_{1_j} * s|[n]; \forall j\}$, are shown in Fig. 4.

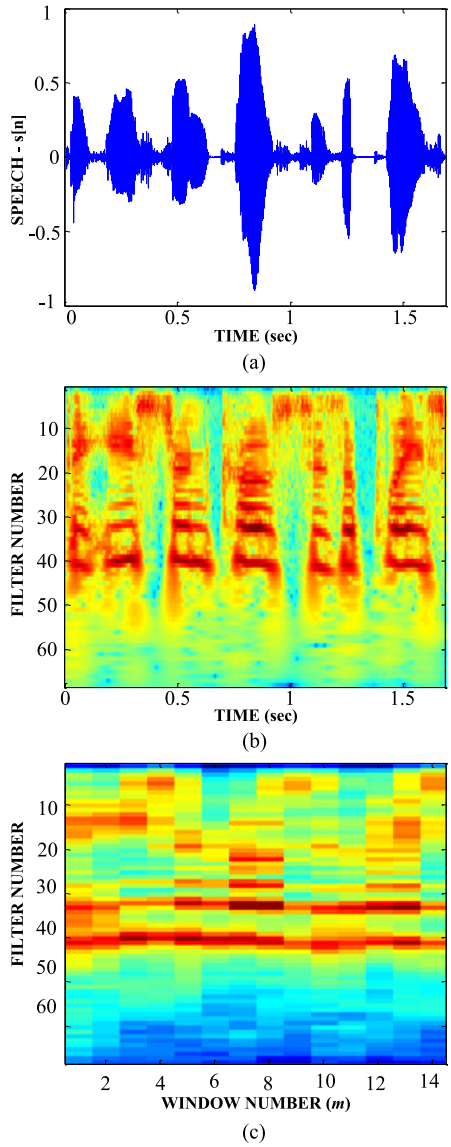

Fig. 4. (a) Speech signal – "The eastern coast is a place", sampled at 16 kHz; (b) log-scalogram of the speech signal; (c) First level log scatter coefficients ($\log S_1[m]$) of the speech signal. Note that the highest filter number corresponds to the lowest centre frequency and $n = mM$ where M denotes window size (M = 4,096).

The second level scattering coefficients $S_2[m]$ are then computed by running the absolute values of the outputs of each of the first level filters through the second level filter-bank and following the same procedure. i.e.,

$$S_2\,[m] = \left\{\left(\phi * \left|\psi_{2_i} * \left|\psi_{1_j} * s\right|\right|\right)[mM];\ \ \forall i, j\right\} \tag{6}$$

where, $i \in 1,\ 2, \ldots, N_2$ denotes the i^{th} filter in the second level filter-bank and $j = 1,\ 2,\ \ldots,\ N_1$ denotes the j^{th} filter in the first level filter-bank.

From (5) it can be seen that the square of the zeroth level scattering coefficient is an approximation of the short-term energy of the speech signal, $s[n]$, and the squares of the first level scattering coefficients are estimates of the energies in the sub-bands defined by the first level constant-Q filter-bank. Filter-bank based front-ends such as CFCCs [42] and CQCCs [43]

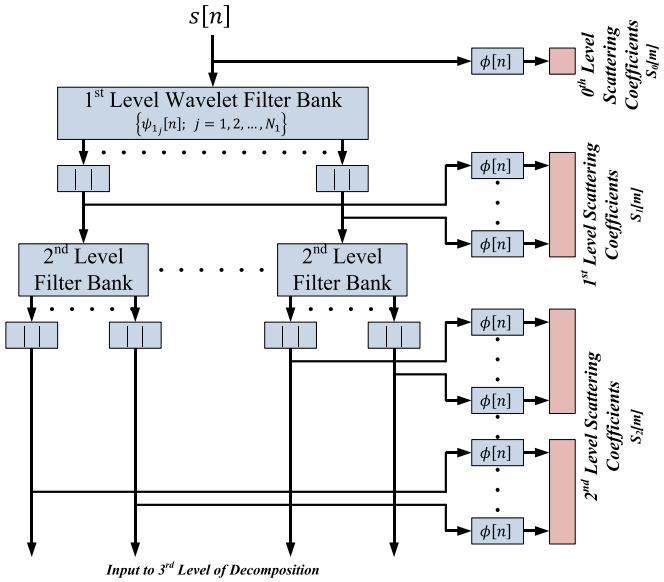

Fig. 5. Two level scattering decomposition.

utilise constant-Q filters that are equivalent to wavelet decompositions (using an appropriate mother wavelet). Consequently the information captured by these front-ends is also captured by log scattering coefficients from the zeroth and first level scattering decomposition. Similarly, MFCCs are also extracted using near constant-Q filters and can be approximated by a suitable wavelet decomposition.

The second level scattering decomposition can be viewed as constant-Q spectral decomposition of instantaneous amplitudes of the spectral components of $s[n]$. In other words, the second level scattering coefficients, $S_2[m]$, are akin to the modulation spectrum of $s[n]$ obtained via constant-Q spectral decomposition. In addition, it should be noted that some information is lost due to the implicit smoothing by $\phi[n]$ when estimating short-term sub-band energy as the first level scattering coefficients (see Fig. 4) and some of this information can be recovered in the second level scattering coefficients. Information lost by the smoothing in the second level scattering can be recovered at the third and subsequent levels, but the amount of useful information relevant to spoofing at these levels is expected to be small and only one or two level scattering coefficients are used in all the experiments reported in this paper. Some information may also be lost when taking the absolute values of the filter outputs at each level prior to smoothing, however, this is not expected to be significant given that it is possible to reconstruct the input audio signals based on the scattering coefficients [53].

Finally it should be noted that the filter with the lowest centre frequency in the wavelet filter-bank has the longest impulse response and the need to constrain this to be shorter than M (the size of the rectangular window, $\phi[n]$, within which the average is computed) imposes a limit on the size of the filter-bank that can be used. This limit in turn leads to the frequency interval below the bandwidth of the lowest frequency filter not being covered by the constant-Q filter-bank. Therefore, a series of $Q - 1$ linearly spaced filters are employed to span this frequency region as in [53]. This addition of linear (not constant-Q) filters to the filter-bank does not alter the scattering decomposition procedure outlined in this section.

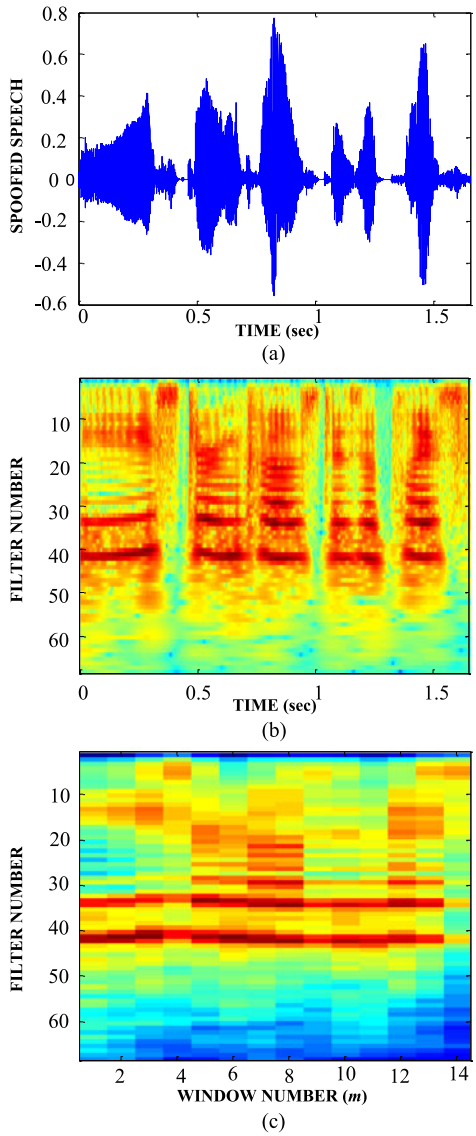

Fig. 6. (a) Spoofed speech signal - "The eastern coast is a place", sampled at 16 kHz; (b) log-scalogram of the spoofed speech signal; (c) First level log scatter coefficients of the spoofed speech signal (window size is 4096 samples).

B. Scattering Spectrum of Spoofed Speech

The log-scalograms (see Section-IV.A) and the first level log-scattering coefficients for a sample of spoofed speech via voice conversion are shown in Fig. 6. It is expected that some of the information that can be used to identify spoofing will be captured by the first-level log-scattering coefficients and the remaining will be captured at higher levels of the scattering decomposition (Fig. 9).

In order to discern the relative discriminative capabilities of the first level scattering coefficients in distinguishing between spoofed and genuine speech, the F-ratio between all genuine speech frames and spoofed speech frames from the training set of the Spoofing and Anti-Spoofing (SAS) corpus [6] (based on

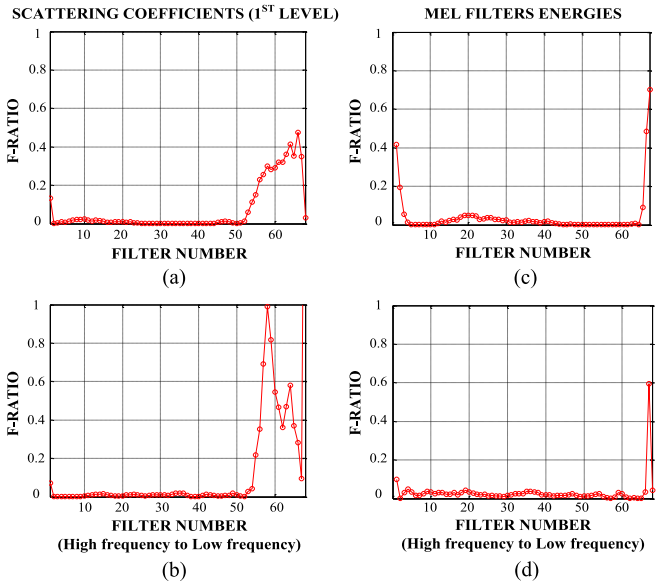

Fig. 7. F-ratios computed for all first level scattering coefficients using 68 filters in the filter-bank (as per Section-V.B) between: (a) genuine speech and spoofed speech obtained via voice conversion; (b) genuine speech and spoofed speech obtained via speech synthesis; and F-ratios computed for mel filter energies using 68 mel-spaced filters between: (c) genuine speech and spoofed speech obtained via voice conversion; (d) genuine speech and spoofed speech obtained via speech synthesis.

all spoofing methods in the database as outlined in Section-V.A) for all coefficients were estimated (Fig. 7(a) and (b)). The F-ratio of the j^{th} scattering coefficient, F_j, is defined as:

$$F_j = \frac{\left(\mu_j^G - \mu_j\right)^2 + \left(\mu_j^S - \mu_j\right)^2}{\left(\sigma_j^G\right)^2 + \left(\sigma_j^S\right)^2} \tag{7}$$

where, μ_j is mean of the j^{th} first level scattering coefficients estimated from all speech frames (spoofed and genuine); μ_j^G and σ_j^G are the mean and standard deviation of the j^{th} first level scattering coefficients estimated from all genuine speech frames; and μ_j^S and σ_j^S are the corresponding mean and standard deviation estimated from spoofed speech.

The F-ratio is the ratio of the separation between the means of the two classes (genuine and spoofed) to the sum of variances of the data from the two classes. A large F-ratio therefore signifies a large separation between relatively tightly clustered sets of points which in turn suggests the two classes are easy to separate.

The plots in Fig. 7(a) and (b) depict F-ratios obtained for the scattering coefficients obtained from the first level decomposition, treating them independently. i.e., joint information between coefficients is not considered. It should be noted that as per (1), the larger the filter number, the lower the centre frequency and smaller the bandwidth. These F-ratios were compared to the F-ratios corresponding to mel filter energies (Fig. 7(c) and (d)). The mel filter energies were extracted using 68 mel-spaced filters (first 60 DCT coefficients were retained) in order to compare with the scattering coefficients that were extracted using 68 filters. The mel filter energies were estimated using 256 ms windows (4,096 samples) to match the window size used in estimating scattering coefficients. It can be seen from the comparison that the scattering coefficients are able to better discriminate between genuine and spoofed speech (both voice conversion and speech synthesis) than mel filter energies.

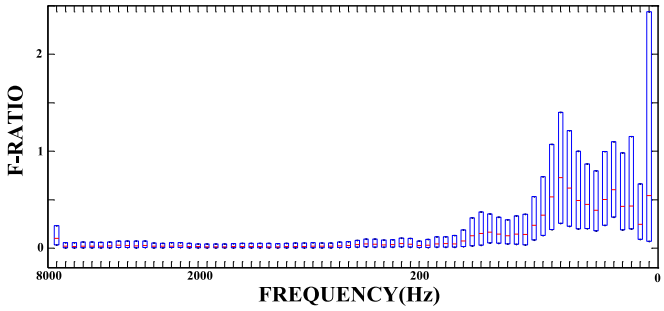

Fig. 8. Boxplots showing median, span from 1st quartile to 3rd quartile of the F-ratios across all spoofed utterances from SAS corpus for 68 first order scattering coefficients.

In order to be doubly sure that the plots in Fig. 7(a) and (b) are general observations across multiple utterances and are not an artefact of just a few frames, all the spoofed utterances in the SAS corpus were compared to genuine speech from the target speaker to estimate F-ratios per utterance for all first level scattering coefficients (68 coefficients). Fig. 8 shows boxplots denoting the median F-ratios across all utterances for each coefficient (red bars) along with the span between 1st and 3rd quartiles (blue boxes).

The plots in Fig. 7(a) and (b) suggest that the higher filters (lower frequency filters) contain most of the information that is useful in discriminating between genuine and spoofed speech. We observe that there is very little discriminative information in the lower indexed filters (higher frequency filters). However, a similar analysis of the F-ratios of the second order scattering coefficients (Fig. 9) revealed that second order coefficients estimated from the outputs of some of these lower indexed filters (higher frequency filters) do contain discriminative information. i.e., even though the first order scattering coefficients (which are conceptually similar to all filter-bank energy based features) are not able to capture information that can help discriminate between genuine and spoofed speech, they are captured at the second level and can potentially lead to better spoofing detection when used in conjunction with the first order coefficients. It can be seen from Figs. 7 and 9 that second order scattering coefficients (corresponding to the 4th to 10th filters) capture discriminative information.

V. Experimental Results

In our experiments, we are particularly interested in comparing the state-of-the-art filter-bank energy based features with those based on the scattering spectrum. Furthermore, we would like to know if there is any additional information obtained from the second level of the hierarchical scattering decomposition. We hoped to further validate the observations made in Section-IV.B through a number of classification experiments. The classification experiments involved the comparison between different front-ends in the context of stand-alone spoofing detection using a single consistent back-end.

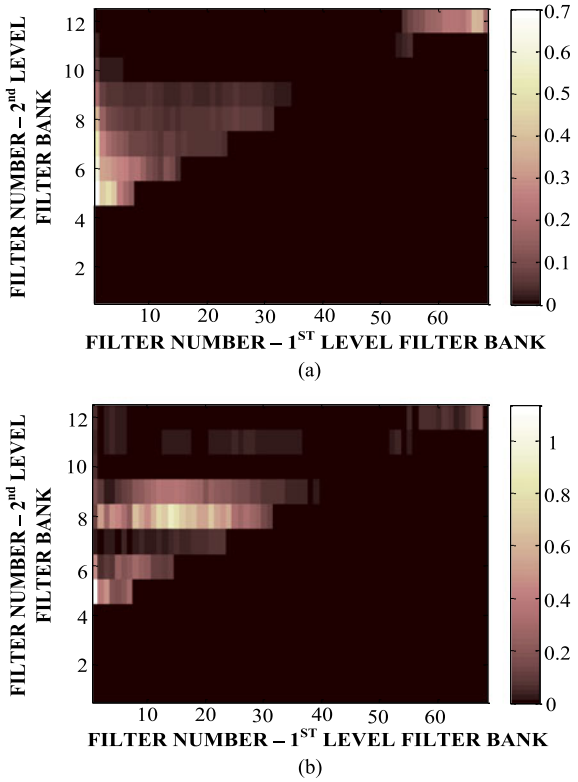

Fig. 9. F-ratio computed for all second level scattering coefficients using 68 filters in the 1st level filter-bank and 12 filters in the 2nd level, between: (a) genuine speech and spoofed speech obtained via voice conversion; (b) genuine speech and spoofed speech obtained via speech synthesis.

A. Database

Countermeasures to spoofing should be agnostic to spoofing methods and be capable of detecting previously unseen spoofing methods. Consequently, in order to analyse and test such countermeasures it is desirable to use a standardised database comprising of multiple spoofing methods encompassing both speech synthesis and voice conversion with established training and test sets. The test set should also include spoofing methods that are not included in the training set in order to simulate unseen spoofing methods. The Spoofing and Anti-Spoofing (SAS) corpus [6] that satisfies all of these requirements are used in all the experiments reported in this paper. The SAS corpus comprises of speech sampled at 16 kHz (recorded at 96 kHz and downsampled) from 45 male and 61 female speakers. This database can be used to evaluate both speaker verification performance (in the presence of spoofing attacks) as well as countermeasure performance (detect spoofed vs genuine speech) and evaluation protocols for both are provided with the database [6]. The training, development and evaluation sets for the countermeasure protocol are summarised in Table I. The performance of standalone countermeasures are evaluated in terms of False Acceptance Rates (FAR) and Equal Error Rates (EER) in this paper.

The spoofed utterances were generated using 13 different spoofing methods of which 5 methods are categorised as known attacks and are used to generate spoofed utterances for the training, development and evaluation sets while the other 8 methods are categorised as unknown attacks and are used to generate spoofed utterances only for the evaluation set. The 5 known attacks in turn comprise of two speech synthesis methods and 3 voice conversion methods while the 8 unknown attacks comprise of 3 speech synthesis methods and 5 voice conversion methods. The 13 attacks are summarised in Table II, details can be found in [6].

TABLE I
SUMMARY OF DATA FOR COUNTERMEASURE PROTOCOL

	No. of speakers		No. of utterances	
	Male	Female	Genuine	Spoofed
Training	10	15	3,750	12,625
Development	15	20	3,497	152,215
Evaluation	20	26	9,404	1,034,397

TABLE II
BRIEF SUMMARY OF SPOOFING ATTACKS IN SAS CORPUS [6]

	Spoofing Technique	Description
Known Attacks	SS-SMALL-16	HMM based speech synthesis using open source "HTS" using few minutes of target speech for adaptation.
	SS-LARGE-16	Same as SS-SMALL-16 but using larger adaptation dataset.
	VC-FEST	Voice conversion using open sourced Festvox system that uses joint density GMMs.
	VC-C1	Voice conversion where only spectral slope was transformed.
	VC-FS	Voice conversion using Frame selection (a simplified version of unit selection).
Unknown Attacks	SS-LARGE-48	Same as SS-LARGE-16 but using 48kHz sampling rate
	SS-SMALL-48	Same as SS-SMALL-16 but using 48kHz sampling rate
	SS-MARY	Speech synthesis using open source Mary-TTS unit selection system [54].
	VC-GMM	Voice conversion based on GMMs (similar to VC-FEST with some enhancement).
	VC-KPLS	Voice conversion using kernel partial least squares regression [55].
	VC-EVC	Voice conversion using eigenvoice GMM [56].
	VC-TVC	Voice conversion using tensor methods [57].
	VC-LSP	Similar to VC-GMM but uses LSP features.

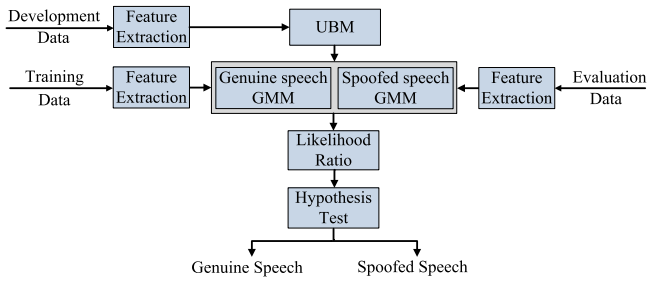

Fig. 10. GMM-UBM back end for spoofing detection system.

TABLE III
NUMBER OF FILTERS AND COEFFICIENTS OF EACH LEVELS WITH $Q_1 = 8$ AND $Q_2 = 1$

M	No. of Filters		No. of Coefficients	
	1st Level	2nd Level	1st Level	2nd Level
1,024	52	10	52	167
4,096	68	12	68	354
16,384	84	14	84	435

TABLE IV
SPOOFING DETECTION PERFORMANCE OF SCCS USING DIFFERENT WINDOW SIZE EVALUATED ON DEVELOPMENT SET OF THE SAS CORPUS [6]

Window Size	M	%EER
64 ms (2^{10} samples)	1,024	0.64
256 ms (2^{12} samples)	4,096	0.31
1,024 ms (2^{14} samples)	16,384	4.19

B. Spoofing Detection System

Without the need to modify an existing speaker verification system, we adopt the stand-alone approach (Fig. 3) in the following experiments. Specifically a spoofing detection system using a GMM-UBM back-end (Fig. 10) is employed.

Front-end: The first and second level scattering coefficients are used as the features in the spoofing detection system and compared to common and established front-ends such as MFCCs, CQCCs (constant-Q cepstral coefficients) [43], CFCCIFs (Cochlear Filter Cepstral Coefficients and Instantaneous Frequency) [42], etc. The Morlet wavelet (other wavelets can also be used) is used for the scattering decomposition at both levels with $Q_1 = 8$ (number of filters per octave at first level) and $Q_2 = 1$ (with $Q_1 = 8$, the Morlet wavelet filters match the frequency resolution of mel filters). These scattering coefficients are computed using overlapping (50% overlap) windows of 256 ms duration ($M = 4{,}096$). In addition, other window sizes of 64 ms ($M = 1{,}024$) and 1,024 ms ($M = 16{,}384$) were also investigated (Table IV). The total number of filters in both the first and second level filter-banks and the number of scattering coefficients at both levels corresponding to all three window sizes are given in Table III.

It should be noted that the number of second level scattering coefficients is less than the number of filters in the first level times the number of filters in the second level (as one would expect) since the fast scattering decomposition outlined in [53] was employed. The scattering coefficients were estimated using a publicly available toolbox [58] http://www.di.ens.fr/data/scattering/. Finally, the feature vector is computed by taking a Discrete Cosine Transform (DCT) of the vector obtained by concatenating the logarithms of the scattering coefficients from all levels as given in (8) and retaining the first 60 of these coefficients. These features are referred to as Scattering Cepstral Coefficients (SCC). Unlike in CQCC, the frequency scale of the scattering coefficients is not re-linearized before applying DCT. A voice activity detector, identical to the one used in [59], was employed to remove unvoiced regions prior to feature extraction.

$$SCC\,[m] = DCT_{60}\left\{[S_0\,[m]\,,\ S_1\,[m]\,,\ S_2\,[m]]\right\} \tag{8}$$

where, DCT_{60} denotes that the first 60 DCT coefficients are retained.

Back-End: A 2-class GMM-UBM back-end (Fig. 10) was used to obtain the log-likelihood ratio between genuine and spoofed speech as below:

$$LLR\,(X) = \log P\,(X|\theta_g) - \log P\,(X|\theta_s) \tag{9}$$

where, X denotes the set of feature vectors from a test utterance, θ_g denotes the Gaussian mixture model (GMM) corresponding to genuine speech and θ_s denotes the GMM corresponding to spoofed speech. Both θ_g and θ_s are estimated via mean only MAP adaptation from a universal background Gaussian mixture model to model the training data. The UBM was trained on the development set (Table I). All the GMMs employed in this system utilised 512 mixture components as in [60]. We also compared the GMM-UBM approach to a GMM approach using maximum likelihood estimation via the EM algorithm. Since the GMM-UBM approach resulted in better classification accuracies, only the results based on the GMM-UBM system are reported in this paper.

C. Performance Metric

Given a stand-alone spoofing detection system, we report two types of errors: (a) False Acceptance Rate (FAR) is the ratio of the number of spoofed utterances accepted as genuine utterances to the total number of trials; and (b) False Rejection Rate (FRR) is the ratio of the number of genuine speech utterances rejected as spoofed utterances to the total number of trials. As indicated in Fig. 10, the spoofing detection system compares a log-likelihood ratio (or equivalent score) against a predetermined threshold in a hypothesis test. The threshold at which the FAR is equal to the FRR is referred to as the equal error rate point, and the FAR/FRR at this point is referred to as the Equal Error Rate. Equal Error Rate is adopted as the performance metric for comparisons of spoofing detection systems in this paper. EER was also adopted as the evaluation metric in the ASVspoof 2015 challenge [61].

TABLE V
COMPARISON OF %EER BETWEEN SCC AND CQCC-A ON THE SAS CORPUS [6] TEST SET

	Types of spoofing	CQCC-A	SCC
Known	VC-C1	0.47	0.05
	VC-FS	0.02	0.02
	VC-FEST	0.28	0.02
	SS-LARGE-16	0.05	0.02
	SS-SMALL-16	0.02	0.02
Unknown	VC-KPLS	0.59	0.13
	VC-GMM	0.54	0.02
	VC-EVC	12.93	0.09
	VC-TVC	11.25	0.26
	VC-LSP	0.49	0.07
	SS-LARGE-48	0.00	0.00
	SS-SMALL-48	0.00	0.00
	SS-MARY	3.93	17.58
	Known	0.23	**0.04**
	Unknown	5.96	**3.96**
	Overall	4.41	**2.92**

D. Experimental Results

Preliminary experiments were carried out on the development set of the SAS corpus to choose a window size from 64 ms, 256 ms, and 1024 ms. It should be noted that the development set only contained the 5 known spoofing attacks and none of the unknown attacks which were only present in the test set (see Table II). The performance corresponding to these window sizes in terms of %EER are reported in Table IV. Based on these results, a window size of 256ms was chosen for all subsequent experiments.

The SCC features were also compared to a baseline system using CQCC-A features and the results are tabulated in Table V. CQCC-A features were chosen as the baseline for this comparison since there are no previously reported directly comparable results on the SAS corpus and CQCC-A features have been shown to outperform all other features on the ASVspoof 2015 challenge corpus [61]. The CQCC-A results given in Table V were estimated using feature extraction code [60] published by the authors of [43] using a window size of 8 ms. Both the SCC and the CQCC-A based systems employed identical back-ends outlined in Fig. 10. In addition to the overall performance on the SAS test set, Table V also provides the EER obtained by both systems in detecting the test utterances corresponding to each of the 13 spoofing attacks in the corpus as well as overall performances on the 5 known and 8 unknown attacks (see Table II for a summary of spoofing attacks in SAS corpus). We observe in Table V that the system based on SCC features outperforms the baseline system in both known and unknown cases with the overall error rate being almost halved.

We do observe an exception with the SS-MARY data. It is interesting to note that SS-MARY is the only unit selection based speech synthesis method in this database. Even though the distortions at unit boundaries can be audible, SS-MARY speech has been found to be hard to detect by most spoofing detection methods reported so far [44]. It is also interesting to note that the CQCC-A features comprise only the acceleration ($\Delta - \Delta$) coefficients. When the static CQCCs are combined with the acceleration coefficients, the EER in detecting SS-MARY increases significantly [43]. This suggests that information that characterises the unit selection spoofing method, such as unit boundaries, is captured better by the acceleration coefficients than by the static coefficients. The proposed SCC features were not implemented as an analogous counterpart to the acceleration coefficients and hence may not match the performance of CQCC-A for the SS-MARY data. However the proposed SCC features outperform CQCC-A when detecting all forms of voice conversion and parametric speech synthesis methods in the database.

TABLE VI
COMPARISON OF SYSTEM PERFORMANCE IN TERMS OF %EER FOR 1ST AND 2ND LEVEL COEFFICIENTS ON SAS CORPUS [6] TEST SET

Levels	Known	Unknown	Overall
1st level SCC	0.55	4.90	3.89
2nd level SCC	0.04	3.96	2.92

TABLE VII
COMPARISON OF SCC TO ALL KNOWN SYSTEMS ON THE ASVSPOOF 2015 CHALLENGE CORPUS [61] IN TERMS OF %EER

System	Known	Unknown	Overall
CFCC-IF [42]	0.41	2.01	1.21
i-vector [39]	0.01	3.92	1.97
M&P features [40]	0.00	5.22	2.61
LFCC-DA [41]	0.11	1.67	0.89
CQCC-A [43]	0.05	0.46	0.26
SCC	**0.02**	**0.33**	**0.18**

In order to quantify the benefit of incorporating information contained in the second level of the scattering decomposition, the system based on SCC obtained with a two level scattering decomposition was compared to a system based only on the first level scattering decomposition on the SAS test corpus (Table VI). The SCC features based only on the first level decomposition are obtained by taking the first 60 DCT coefficients of only the first level scattering coefficients. It can be seen that the addition of information from second level scattering coefficients substantially improves performance.

Finally, the spoofing detection system using SCCs was also evaluated on the ASVspoof 2015 challenge corpus [61] and results are compared to some best previously reported results on this database known to the authors in Table VII. It can be seen that the system based on SCCs, outperformed all previously reported systems on this database. Note that a 2-class GMM back-end (trained via Maximum Likelihood estimation) was utilised since the systems, reported in Table VII, were all evaluated with this back-end. The performances of the previously reported CQCC-A system [43] and the SCC based system for the individual known and unknown spoofing attacks contained in the ASVspoof 2015 challenge corpus are provided in Table VIII.

TABLE VIII
COMPARISON OF %EER BETWEEN SCC AND CQCC-A ON THE ASVSPOOF 2015 CHALLENGE CORPUS [61] TEST SET

	Types of spoofing	CQCC-A [43]	SCC
Known	VC-FS	0.01	0.01
	VC-C1	0.11	0.12
	SS-LARGE-16	0.00	0.00
	SS-SMALL-16	0.00	0.00
	VC-FEST	0.13	0.02
Unknown	VC-GMM	0.10	0.01
	VC-LSP	0.06	0.01
	VC-TVC	1.03	0.03
	VC-KPLS	0.05	0.01
	SS-MARY	1.07	3.94
	Known	0.05	**0.02**
	Unknown	0.46	**0.33**
	Overall	0.26	**0.18**

VI. CONCLUSION

This paper proposes hierarchical scattering decomposition as a general framework that can be used to design effective front-ends for stand-alone spoofing detection. The first level scattering coefficients provide a constant-Q spectral decomposition that is as competitive as other filter-bank front-ends. In addition, the availability of the second (and higher) level of decomposition allows for information lost in the previous level to be recovered. Experimental results in this paper demonstrate that the performance of features based on the first level of the scattering decomposition matches that of the best front-ends previously reported in literature and the addition of information from the second level of decomposition improves performance even further. It is expected that the hierarchical filter-bank structure of the scattering decomposition negates the need for using velocity and acceleration coefficients as additional features. This is supported by the experimental results reported in this paper where Scattering Cepstral Coefficients (SCCs) based on a two-level decomposition outperforms filter-bank features that utilise velocity and acceleration coefficients. It should be noted the SCCs provide the best performance among all systems benchmarked on the ASVspoof 2015 challenge corpus till date (Table VII). Finally, SCC features are also compared to CQCC-A features, a state-of-the-art front-end on ASVspoof task, on the larger SAS corpus and shown to be superior. Future work will look at alternative feature representations that can more effectively use information from higher levels of scattering decomposition.

REFERENCES

[1] K. A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," *IEEE Signal Process. Soc. Speech Lang. Tech. Committee Newslett.*, 2013. [Online]. Available: http://archive.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2013-02/SpeakerVerificationMakesItsDebutinSmartphone/

[2] Nuance, "Nuance vocalpassword," 2015. [Online]. Available: http://www.nuance.com/for-business/customer-service-solutions/voice-biometrics/vocalpassword/index.htm

[3] USAA, "Easily and Securely log on to the USAA app," 2015. [Online]. Available: https://www.usaa.com/inet/pages/enterprise_howto_biometrics_landing_mkt?akredirect = true

[4] D. Mukhopadhyay, M. Shirvanian, and N. Saxena, "All your voices are belong to us: Stealing voices to fool humans and machines," in *Proc. Eur. Symp. Res. Comput. Security*, 2015, pp. 599–621.

[5] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. De Leon, "Speaker recognition anti-spoofing," in *Handbook of Biometric Anti-Spoofing*. Berlin, Germany: Springer, 2014, pp. 125–146.

[6] Z. Wu *et al.*, "SAS: A speaker verification spoofing database containing diverse attacks," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4440–4444.

[7] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. 2004 Int. Symp. Intell. Multimedia, Video Speech Process.*, 2004, pp. 145–148.

[8] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification-a study of technical impostor techniques," in *Proc. EUROSPEECH*, 1999, pp. 1211–1214.

[9] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 8, pp. 2280–2290, Oct. 2012.

[10] 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. 2012 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 4401–4404.

[11] 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. 2012 Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf.*, 2012, pp. 1–5.

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

[13] S.-H. Chen and Y.-R. Luo, "Speaker verification using MFCC and support vector machine," in *Proc. Int. MultiConference Eng. Comput. Scientists*, 2009, pp. 18–20.

[14] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," *Speech Commun.*, vol. 55, pp. 237–251, 2013.

[15] S. H. Ghalehjegh and R. C. Rose, "Deep bottleneck features for i-vector based text-independent speaker verification," in *Proc. 2015 IEEE Workshop Autom. Speech Recog. Understanding*, 2015, pp. 555–560.

[16] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah, "A model based voice activity detector for noisy environments," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2297–2301.

[17] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," in *Proc. INTERSPEECH*, 2015, pp. 1146–1150.

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

[19] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. 2014 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1695–1699.

[20] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey, Speaker Lang. Recognit. Workshop*, 2010, p. 14.

[21] H. Zen *et al.*, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. 6th ISCA Workshop Speech Synthesis*, 2007, pp. 294–299.

[22] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proc. INTERSPEECH*, 2015, pp. 879–883.

[23] T. Merritt, R. A. Clark, Z. Wu, J. Yamagishi, and S. King, "Deep neural network-guided unit selection synthesis," in *Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5145–5149.

[24] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7962–7966.

[25] G. Coorman, J. Fackrell, P. Rutten, and B. Van Coile, "Segment selection in the L&H Realspeak laboratory TTS system," in *Proc. INTERSPEECH*, 2000, pp. 395–398.

[26] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. 1996 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996, pp. 373–376.
[27] H. Zen, "Statistical parametric speech synthesis," UK Speech Conf., Edinburgh, UK, 2014.
[28] Y. Stylianou, "Voice transformation: A survey," in *Proc. 2009 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 3585–3588.
[29] N. Evans, F. Alegre, Z. Wu, and T. Kinnunen, "Anti-spoofing, voice conversion," in *Encyclopedia of Biometrics*. Berlin, Germany: Springer, 2015, pp. 115–122.
[30] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.
[31] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. 1988 Int. Conf. Acoust., Speech, Signal Process.*, 1988, pp. 655–658.
[32] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. 2000 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, pp. 1843–1846.
[33] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Proc. INTERSPEECH*, 2011, pp. 653–656.
[34] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Proc. 2001 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 841–844.
[35] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 922–931, Jul. 2010.
[36] 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, pp. 61–65.
[37] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah, "Investigation of sub-band discriminative information between spoofed and genuine speech," in *Proc. INTERSPEECH*, San Francisco, CA, USA, 2016, pp. 1710–1714.
[38] L.-W. Chen, W. Guo, and L.-R. Dai, "Speaker verification against synthetic speech," in *Proc. 2010 7th Int. Symp. Chin. Spoken Lang. Process.*, 2010, pp. 309–312.
[39] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge," in *Proc. 2016 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5475–5479.
[40] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, 2015, pp. 2052–2056.
[41] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2087–2091.
[42] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. INTERSPEECH*, 2015, pp. 2062–2066.
[43] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. ODYSSEY, Speaker Lang. Recog. Workshop*, Bilbao, Spain, 2016.
[44] Z. Wu *et al.*, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 768–783, Apr. 2016.
[45] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2082–2086.
[46] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Proc. INTERSPEECH*, 2015, pp. 2077–2081.
[47] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2072–2076.
[48] P. Taylor, *Text-to-Speech Synthesis*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[49] N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu, "Robust deep feature for spoofing detection-the SJTU system for ASVspoof 2015 challenge," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2097–2101.
[50] M. J. Alam, P. Kenny, V. Gupta, and T. Stafylakis, "Spoofing detection on the ASVspoof2015 challenge corpus employing deep neural networks," in *Proc. ODYSSEY, Speaker Lang. Recognit. Workshop*, 2016, pp. 270–276.
[51] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 16, pp. 1791–1801, Aug. 2011.
[52] S. Mallat, "Group invariant scattering," *Commun. Pure Appl. Math.*, vol. 65, pp. 1331–1398, 2012.
[53] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.
[54] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *Int. J. Speech Technol.*, vol. 6, pp. 365–377, 2003.
[55] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 806–817, Mar. 2012.
[56] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Non-parallel training for many-to-many eigenvoice conversion," in *Proc. 2010 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4822–4825.
[57] D. Saito, N. Minematsu, and K. Hirose, "Effects of speaker adaptive training on tensor-based arbitrary speaker conversion," in *Proc. INTERSPEECH*, 2012, pp. 98–101.
[58] Scattering. 2016. [Online]. Available: http://www.di.ens.fr/data/scattering/
[59] M. Nosratighods, T. Thiruvaran, J. Epps, E. Ambikairajah, M. Bin, and L. Haizhou, "Evaluation of a fused FM and cepstral-based speaker recognition system on the NIST 2008 SRE," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 4233–4236.
[60] M. Todisco, H. Delgado, and N. Evans, "CQCC features for spoofed speech detection," 2016. [Online]. Available: http://audio.eurecom.fr/content/software
[61] Z. Wu *et al.*, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH*, 2015, pp. 2037–2041.

Kaavya Sriskandaraja received the B.Sc.(Hons.) degree in engineering from the University of Peradeniya, Peradeniya, Sri Lanka, in 2012. She is currently working toward the Ph.D. degree with the Speech Processing Research Group in the School of Electrical Engineering and Telecommunications, University of New South Wales, N.S.W., Australia. Her research interests include speaker verification, spoofing and anti-spoofing, and machine learning. She is a Member of ISCA.

Vidhyasaharan Sethu received the B.E. degree from Anna University, Chennai, India, in 2005, and the M.Eng.Sc. degree in signal processing and the Ph.D. degree from the University of New South Wales (UNSW), N.S.W., Australia, in 2006 and 2010, respectively. He is currently a Lecturer at the School of Electrical Engineering and Telecommunications, UNSW. From 2010 to 2013, he was a Postdoctoral Fellow at the Speech Processing Research Group, UNSW. His research interests include the application of machine learning to speech processing and computational paralinguistic, speaker recognition, and language identification. He has authored and co-authored approximate 40 conferences and journal papers. He is a Reviewer for IEEE, IET, and EURASIP journals. Dr. Sethu is a Member of ISCA and APSIPA.

Eliathamby Ambikairajah received the B.Sc. (Hons.) degree in engineering from the University of Sri Lanka, , Moratuwa, Sri Lanka and the Ph.D. degree in signal processing from Keele University, Keele, U.K. He was appointed as the Head of Electronic Engineering and later Dean of Engineering at the Athlone Institute of Technology, Republic of Ireland from 1982 to 1999. His key publications led to his repeated appointment as a short-term Invited Research Fellow with the British Telecom Laboratories, U.K., for ten years from 1989 to 1999. He is currently the Head of School of Electrical Engineering and Telecommunications, University of New South Wales (UNSW), N.S.W., Australia. His research interests include speaker and language recognition, emotion detection, and biomedical signal processing. He has authored and co-authored approximately 300 journals and conference papers and is the recipient of many competitive research grants. He is also a regular Reviewer for several IEEE, IET, and other journals and conferences. For his contributions to speaker recognition research, he was invited as a Visiting Scientist to the Institute of Infocomms Research (A*STAR), Singapore in 2009, where he is currently a Faculty Associate. He received the Vice-Chancellor's Award for Teaching Excellence in 2004 for his innovative use of educational technology, the School Awards for Teaching Excellence in 2003, Academic Management in 2001, and the UNSW Excellence in Senior Leadership Award in 2014. Prof. Ambikairajah was an APSIPA Distinguished Lecturer for the 2013–2014 term. He is a Fellow and a Chartered Engineer of the IET UK and Engineers Australia and is a Member of the ISCA and APSIPA.

Haizhou Li (M'91–SM'01–F'14) received the B.Sc., M.Sc., and Ph.D degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively. He is currently a Professor at the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and a Conjoint Professor of the University of New South Wales, N.S.W., Australia. His research interests include automatic speech recognition, speaker and language recognition, natural language processing, and computational intelligence. He has taught in The University of Hong Kong (1988–1989), South China University of Technology, Guangzhou, China (1990–1994), and Nanyang Technological University, Singapore (2006–2016). He was a Visiting Professor at the CRIN/INRIA, France (1994–1995), Research Manager in Apple-ISS Research Centre (1996–1998), Research Director of Lernout & Hauspie Asia Pacific (1999–2001), Vice President of InfoTalk Corp. Ltd (2001–2003), Principal Scientist and Department Head of Human Language Technology, Institute for Infocomm Research, Singapore (2003–2016). Prof. Li is currently the Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2015–2017), a Member of the Editorial Board of Computer Speech and Language (2012–2017), the President of the International Speech Communication Association (2015–2017), and the President of Asia Pacific Signal and Information Processing Association (2015–2016). He was an Elected Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the General Chair of ACL 2012, and INTERSPEECH 2014. He recieved the National Infocomm Award 2002 and the President's Technology Award 2013, Singapore. He was named one of the two Nokia Visiting Professors by the Nokia Foundation in 2009.