# Significance of Source-Filter Interaction for Classification of Natural *vs.* Spoofed Speech

Tanvina B. Patel, *Student Member, IEEE*, and Hemant A. Patil, *Member, IEEE*

***Abstract*—Countermeasures used to detect synthetic and voice-converted spoofed speech are usually based on excitation source or system features. However, in the natural speech production mechanism, there exists nonlinear Source-Filter (S-F) interaction as well. This interaction is an attribute of natural speech and is rarely present in synthetic or voice-converted speech. Therefore, we propose features based on the S-F interaction for a Spoofed Speech Detection (SSD) task. To that effect, we estimate the voice excitation source (i.e., differenced glottal flow waveform, $\dot{g}(t)$) and model it using the well-known Liljencrants-Fant (LF) model to get coarse structure, $g_c(t)$. The residue or difference, $g_r(t)$, between $\dot{g}(t)$ and $g_c(t)$ is known to capture the nonlinear S-F interaction. In the time domain, the $L^2$ norm of $g_r(t)$ in the closed, open and return phases of the glottis are considered as features. In the frequency domain, the Mel representation of $g_r(t)$ showed significant contribution in the SSD task. The proposed features are evaluated on the first ASVspoof2015 challenge database using a Gaussian Mixture Model (GMM)-based classification system. On the evaluation set, for vocoder-based spoofs (i.e., S1-S9), the score-level fusion of residual energy features, Mel representation of the residual signal and Mel Frequency Cepstral Coefficients (MFCC) features gave an Equal Error Rate (EER) of *0.017%* which is much less than the *0.319%* obtained with MFCC alone. Furthermore, the residues of the spectrogram (as well as the Mel-warped spectrogram) of estimated $\dot{g}(t)$ and $g_c(t)$ are also explored as features for the SSD task. The features are evaluated for robustness in the presence of additive white, babble and car noise at various Signal-to-Noise Ratio (SNR) levels on the ASVspoof2015 database and for channel mismatch condition on the Blizzard Challenge 2012 dataset. For both cases, the proposed features gave significantly less EER than that obtained by MFCC on the evaluation set.**

***Index Terms* — Anti-spoofing, source-filter interaction, LF-model, residue, Gaussian mixture model.**

## I. Introduction

THE problem of detecting spoofed speech has been of wide research interest recently. A detailed description of previous studies on the effect of various spoofing attacks on Automatic Speaker Verification (ASV) systems is presented in [1]. An anti-spoofing task finds its application in safeguarding ASV systems against threats to spoofing attacks. It is known that apart from replay and mimicry, ASV systems are highly vulnerable to speech synthesis and voice conversion attacks. Both Synthetic Speech (SS) and Voice-Converted (VC) speech are known to degrade the performance of the current state-of-the-art Joint Factor Analysis (JFA) [2] and Probabilistic Linear Discriminant Analysis (PLDA)-based ASV systems [3]. It has been investigated in [1] that SS generated by state-of-the-art Hidden Markov Model (HMM)-based Text-to-Speech (TTS) synthesis systems (HTS) [4], [5], [6] and VC speech [7]- [8] can degrade the performance of ASV systems. To encourage future research in anti-spoofing, the first ASV spoof 2015 challenge was organized as a special session of INTERSPEECH 2015 that used the Spoofing Anti-Spoofing (SAS) database [9] and provided a common platform for evaluating countermeasures [10]. Several countermeasures were proposed at the challenge to detect SS and VC speech spoof. For the ASV spoof 2015 challenge, the authors proposed a Cochlear Filter Cepstral Coefficients and Instantaneous Frequency (CFCCIF) feature set, which was found to be relatively the best performing system [11].

Most of the countermeasures proposed at the challenge were phase-based (due to the fact that vocoders used in SS and VC speech generation process lack vital phase information, which is also significant for various speech processing tasks including speech perception and synthesis). However, phase-based countermeasures are not effective for vocoder-independent spoofing techniques such as Modular Architecture for Research on speech sYnthesis (MARY) TTS systems [12], [13]. With the possibility of using a phase-based vocoder in speech synthesis techniques, the phase-based countermeasures alone may not be effective [14]. Thus, other research directions, such as the use of excitation source-based information and the nonlinearity in the human speech production mechanism can be explored for the Spoofed Speech Detection (SSD) task. Earlier studies using source-based features include the use of the Fundamental Frequency ($F_0$) [15], the $F_0$ contour [16] and its variability [17] to detect SS spoof. Very recently, it was proposed to use $F_0$ and its correlation with the Strength of Excitation (SoE) estimated from speech and voice excitation source for the SSD task [18]. In addition, features such as the Linear Prediction (LP), Linear Term Prediction (LTP) and Non-Linear Prediction (NLP) residual have been used for the SSD task [19], [20]. Various other countermeasures have been proposed such as Modified

This work was supported in part by the Department of Electronics and Information Technology (DeitY), Government of India, through two consortium projects, TTS Phase-II and ASR Phase-II, and in part by the authorities of DA-IICT, Gandhinagar, India.

Tanvina B. Patel and Hemant A. Patil are with the Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar 382007, Gujarat, India (e-mail: tanvina_bhupendrabhai_patel@daiict.ac.in; hemant_patil@daiict.ac.in).

Group Delay (MGD)-based features [21], [22] subband processing based Linear Frequency Cepstral Coefficients (LFCC) [23], [24] and Constant Q Cepstral Coefficients (CQCC) features [25]. In addition, the use of Deep Neural Networks (DNN)-based representation [26], [27], [28], i.e., the use of deep features has also been explored for the SSD task.

The vocal folds along with their dynamic movements represent various aspects of both speech and the speaker. During phonation, the gradual opening of the glottis and its sudden closure result in an *asymmetric* shape of the glottal flow waveform. Assuming a Linear Time-Invariant (LTI) speech production mechanism, the derivative (due to lip radiation [29], [30]) of this glottal flow waveform is referred to as the voice excitation source (i.e., $\dot{g}(t)$). This excitation source can be parameterized using physical or acoustical models. Physical models such as the two-mass model of Ishizaka and Flanagan involve the use of a large number of independent parameters for modeling [31]. The acoustic Liljencrants-Fant (LF) model is also a good approximation to $\dot{g}(t)$ and it can be represented in the frequency domain as well [32]. The LF-model gives the shape and timing parameters of the voice excitation source that are known to relate to voice quality measures, such as Speed Quotient (SQ), Open Quotient (OQ) and Return Quotient (RQ) [33]. Natural speech has variations ranging from creaky to breathy voice, which need to be incorporated in synthesis and voice conversion techniques for better speech quality. In this context, incorporating the excitation source information through the LF-model into an HMM-based synthesizer has shown to give more naturalness and reproduction of basic voice qualities such as breathy and tense [34]. Another example is of GlotHMM, which uses inverse filtering for generating glottal excitation and modeling it into an HMM framework using Line Spectral Frequencies (LSFs) [35]. Other earlier source models in parametric speech synthesizers include a simple pulse/noise excitation model [36], Multi-Band mixed Excitation (MBE) [37], Speech Transformation and Representation using Adaptive Interpolation weiGHTed spectrum (STRAIGHT) vocoder that uses a mixed excitation model [38] and the Harmonic-plus-Noise Model (HNM) of speech [39]. The use of system-level features to model the vocal tract filter is very well-known in speech synthesis and voice conversion techniques. These system-level features include the Mel Frequency Cepstral Coefficients (MFCC) [7], generalized Mel-cepstral Coefficients (MCC) [40]- [41], LSF representation of Linear Prediction Coefficients (LPC) [42] and approaches such as STRAIGHT-based speech parameters encoded into MCCs or LSFs. Thus, the majority of the techniques exist to model either the source or system characteristics individually. However, it is not only the independent role of the source or system that contributes in producing natural speech; rather it is also the time-varying dependencies between them or the nonlinear source-filter (S-F) *interaction* that effectively contributes to naturalness and speaker identity [43]- [44].

## II. Basis of the Proposed Approach

In the linear S-F theory, the source of speech production is independent of the vocal tract (filter). In such a case, the source impedance is much higher than the input impedance to the vocal tract. However, due to the narrow constriction of the vocal tract above the glottis, there exists a nonlinear S-F interaction. The closer the constriction is to the vocal folds, greater is the degree of interaction [45]. In this case, the source impedance is comparable to the vocal tract input impedance. This makes the glottal flow highly dependent on the acoustic pressures in the vocal tract. According to the landmark investigations in [45], [46], [47], there are two primary levels of S-F interaction, namely, Level 1 and Level 2. The Level 1 interaction occurs due to feedback from the vocal tract acoustic pressure (i.e., standing waves) that imparts variations in the glottal airflow (i.e., transglottal pressure drives the glottal flow which in turn is affected by the epiglottis pressure). On the other hand, Level 2 interaction primarily occurs in cases with high $F_0$ where the pitch harmonics are near the formants. It is responsible for variations in vocal fold vibrations (i.e., tissue movements) that occur with same pressure (i.e., intraglottal pressure drives the vocal folds) [45].

The findings in [45] summarize that there always exists an interaction of glottal airflow with the acoustic vocal tract. The primary effect of Level 1 interaction is the glottal airflow skewing (which in turn, balances source spectrum in terms of odd and even harmonics) that can be expressed by an analytic formula [32], [35]. The pressure from the vocal tract against the glottis will slow the flow and change its skewness. In addition to the asymmetric glottal flow, simulation of a simplified electrical first formant model showed the presence of a sinusoidal ‘ripple’ component (i.e., a fine structure superimposed onto the coarse structure) onto the open phase of the glottal flow [29]. Other effects of S-F interaction include abrupt increase in the first formant ($F_1$) and the corresponding *-3* dB bandwidth when the glottis opens. The increase in the bandwidth of the formant causes sudden decay of the vocal tract impulse response within a glottal cycle and is responsible for truncation effect in the speech waveform [29].

In [48], a structure for $\dot{g}(t)$ that includes both coarse structure and ripple component (produced as a result of S-F interaction) was proposed. The use of ripple (i.e., the fine structure features) is explored in the speaker identification task [43], [49]. However, its use in speech synthesis or voice conversion was not significantly explored. Thus, SS or VC speech may not sound as natural and intelligible to a certain extent due to the lack of nonlinear S-F interaction information. To that effect, in this study, we explore the fact that the nonlinear S-F interaction is an attribute of the natural speech production mechanism and not that of machine-generated speech. It is highly complex for speech synthesis and voice conversion techniques to build or mimic such S-F interaction. With this motivation, we study the differences between the actual voice excitation source $\dot{g}(t)$ and its coarse structure, $g_c(t)$ (i.e., fitted LF-model). The voice excitation source $\dot{g}(t)$ is estimated by using a linear inverse filtering technique. The use

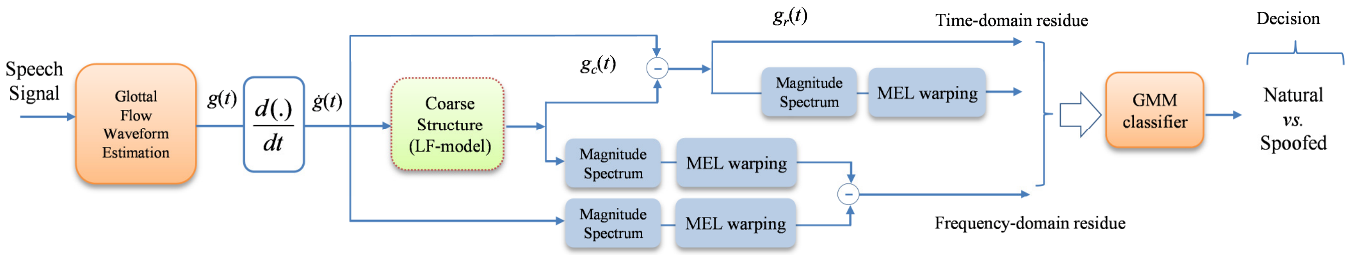

Fig. 1: Proposed approach to use the glottal source features (both in time and frequency domains) for the Spoofed Speech Detection (SSD) task

of a linear inverse filtering does not capture the effects of time variance in formant frequencies (due to the inherent segment-level block-based processing in the LP approach). That is, the true formants changes when the vocal folds are open *vs.* when they are closed are not captured. However, the anti-resonances of the glottal inverse filtering method remain fixed over many pitch periods. Thus, on using a linear fit to model the glottal flow, the anti-resonances that appear as time-varying amplitude variations in the open phase of the glottal flow are captured as ripple structures.

To fit the acoustic LF-model, an exhaustive search method is used to obtain $g_c(t)$ for an estimate of $\dot{g}(t)$ [50]. This approach varies the shape parameter $R_d$ within a specific range and attains the best fit depending on the minimum cost obtained both in time and frequency domains. This motivates us to consider the residuals both in the time and frequency domains. The fitted LF-model $g_c(t)$, when subtracted from the $\dot{g}(t)$, gives the residual signal $g_r(t)$. The residual signal has information about the ripple (due to first formant ($F_1$) modulation of the vocal tract system) and the aspiration components (due to turbulence at the vocal folds) [48]. Thus, in the time domain, the $L^2$ norm of $g_r(t)$ in the closed phase, open phase and return phase of the glottis is considered as feature representations for the SSD task. In addition, as the $F_1$ modulation information is in the lower frequency range (*<1 kHz*), we also consider the Mel representation (having high resolution for frequencies *<1 kHz*) of the residual $g_r(t)$. Furthermore, the residual information is also obtained in the frequency domain by using the difference between the spectrogram of $\dot{g}(t)$ and the spectrogram of $g_c(t)$. The residue in the frequency domain is also obtained by using the difference of the Mel warped spectrograms of $\dot{g}(t)$ and $g_c(t)$. Thus, shape and energy-based features in the time domain and several feature representations in the frequency domain are used for the SSD task. The schematic of the best representative features in time and frequency domains are shown in Fig. 1.

On the ASVspoof 2015 challenge database, the proposed features work well for both known and unknown vocoder-dependent attacks. The performance of the vocoder-independent attack was not much improved as these are generated by concatenating natural speech sound units. Hence, the S-F interaction in such a case is similar as in natural speech except at the point of concatenation of speech sound units. The energy-based features had significant contribution in spoof detection even with fewer feature dimensions and further the use of frequency-domain features was found to add complementary information for the SSD task. The ASV spoof 2015 challenge database considers the case of clean speech without additive noise or channel mismatch conditions [10]. Therefore, few studies have recently investigated the performance of the countermeasures in the presence of noise [51] and under channel-mismatch cases [52]. Along the similar lines, we evaluate the features in the presence of additive white, babble and car noise at various Signal-to-Noise Ratio (SNR) levels. In addition to the clean speech case, the proposed time-domain and frequency-domain representations were found to perform well even for all the three cases of noisy signals. Next, to consider channel variability effects, the features are evaluated on the Blizzard Challenge systems [53]. The Blizzard Challenge 2012 database [54], consisting of both Statistical Parametric Speech Synthesis (SPSS) and Unit Selection Synthesis (USS)-based speech is used. It was observed that with the proposed features, the performance was better than MFCC, signifying robustness even with different recording conditions and completely unknown attacks.

## III. Voice Source Parameterization

This section describes the inverse filtering approach to obtain $\dot{g}(t)$, followed by the description of the LF-model and its estimation from the $R_d$ parameter using a search algorithm.

### A. *The Coarse Structure (LF-Model)*

To obtain an initial estimate of voice excitation source $\dot{g}(t)$, we assume the speech production mechanism to be an LTI system, i.e.,

$$s(t) \approx A\frac{d}{dt}\left[g(t)*h(t)\right] = A\left[\frac{d}{dt}g(t)\right]*h(t) = A\dot{g}(t)*h(t), \qquad (1)$$

where * is the convolution operation, $A$ is the gain that controls loudness, $s(t)$ is the speech signal, $\dot{g}(t)$ is derivative of the glottal flow waveform ($g(t)$) and $h(t)$ is the impulse response of the vocal tract system [29]. Thus, to obtain an initial estimate of the $g(t)$, the Iterative Adaptive Inverse Filtering (IAIF) method is used to inverse filter the vocal tract information from the speech signal [55].

The coarse structure $g_c(t)$ is a parameterization of the shape of the excitation source $\dot{g}(t)$, as shown in Fig. 2. According to the time intervals, $g(t)$ is divided into closed phase, open phase and return phase. The shape and flow of the regions represent different attributes of the speaker and the nature of the speech signal. The LF-model can be defined by the five time instants, i.e., the glottal opening time ($t_o$), the instant when the glottis closes ($t_c$)**,** the time when $\dot{g}(t)$ crosses zero ($t_p$), the time when the $\dot{g}(t)$ reaches its maximum negative value ($t_e$) and the time

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2682788, IEEE Journal of Selected Topics in Signal Processing

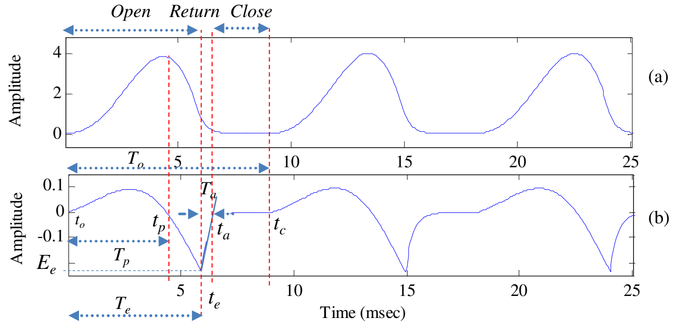

Fig. 2. (a) A schematic of $g(t)$ and (b) the corresponding derivative of the $g(t)$ along with various timing instants and the time periods used in the LF-model.

when tangent to the return phase that crosses the time-axis ($t_a$). Similarly, we can define time periods corresponding to the LF-model as $T_o$ (i.e., duration of a glottal cycle), $T_p$ (period from $t_o$ to $t_p$), $T_e$ (period from $t_o$ to $t_e$) and $T_a$ (period from $t_e$ to $t_a$). Thus, over a glottal cycle, the LF-model consists of an exponentially increasing sinewave, a decaying exponential function, and completion with a zero amplitude region, as described by the following equations [32], [34];

$$g_c(t)\,|_{LF} = \begin{cases} e_{open}(t) = E_0 e^{\alpha t}\sin(\omega_o t), & t_o \le t \le t_e, \\ e_{return}(t) = E_1[e^{-\beta(t-t_e)} - e^{-\beta(t_c-t_e)}], & t_e < t \le t_c, \\ e_{close}(t) = 0, & t_c < t \le T_o, \end{cases} \quad (2)$$

where $E_0 = -E_e/(\sin(\omega_o t_e)e^{\alpha t_e})$ and $E_1 = -E_e/(1-e^{-\beta}(t_c-t_e))$ in eq. (2), parameters $\omega_o$ (angular frequency of the sinewave related to the rise time of the $\dot{g}(t)$) and $\alpha$ (growth factor) determine the shape parameters in the open phase. The parameter $E_e$ (amplitude of maximum excitation) and $\beta$ (exponential time constant) constitute the shape parameters in the return phase. The time instant $t_o$ is assumed to be zero and it is omitted in the formulas. The parameters of the LF-model can be obtained by considering the following assumption [32]:

$$\int_0^{T_0} g_c(t)dt = 0, \quad \therefore e_{open}(t_e) = e_{return}(t_e) = -E_e. \quad (3)$$

In [33], a set of dimensionless parameters of the LF-model, called $R$-parameters, have been derived. These parameters can be expressed as dimensionless quotients often used to describe the shape of the glottal source signal, $\dot{g}(t)$. The $R$-parameters affect the coarse structure representation both in time and frequency domains. The $R$-parameters are given by,

$$R_g = \frac{T_o}{2T_p}\ , \ R_k = \frac{t_e - t_p}{T_p}, \ \text{and} \ \ R_a = \frac{T_a}{T_o}. \quad (4)$$

The parameters $R_k$ and $R_a$ are known to relate to the speed quotient ($SQ$) and return quotient ($RQ$), respectively. The $R$-parameters are related to the open quotient ($OQ$) as [33]:

$$OQ = \frac{1+R_k}{2R_g} + R_a. \quad (5)$$

In [33], an $R_d$ parameter was developed that captures almost all possible variations of the LF-model. The $R_d$ parameter is related to $R_g$, $R_k$, and $R_a$ parameters by the following [33]:

$$R_d = \frac{1}{0.11}\left(0.5 + 1.2R_k\right)\left(\frac{R_k}{4R_g + R_a}\right). \quad (6)$$

There have been various approaches in the literature to determine the coarse structure of the estimated $\dot{g}(t)$. One of the approaches includes minimizing the least square error when $g_c(t)$ in eq. (2) is fitted to the estimate of $\dot{g}(t)$. The error function in such a case is a nonlinear function of the model parameters and needs to be solved iteratively using a nonlinear least squares algorithm [44]. However, this approach depends on the initial estimates of the time and shape parameters. In addition, it is rather difficult to obtain accurately these timing parameters from the speech waveform. Thus, in this work, instead of using an iterative algorithm that optimizes all shape and time parameters, we consider the work carried out in [50], where only the $R_d$ parameter is varied. This approach not only aims at minimizing the error in the time domain but it minimizes the error in the frequency domain as well.

### *1) Determination of GCI and $F_0$*

In this work, we need to estimate the Glottal Closure Instants (GCIs) to fit the LF-model at each glottal cycle in the time domain. Hence, we adapt the time-domain approach to estimate the GCI. The Zero Frequency (ZF) filtering method, also known as the *0-Hz* resonator, is used here [56]. The basic idea behind the ZF method is that the effect due to an impulse is spread uniformly across *all* the frequency regions including zero frequency. Thus, by passing the speech signal through the ZF filter, we decouple the interference of the vocal tract system (whose resonances are at much higher frequencies than zero frequency) from the excitation source. Therefore, the speech signal is passed through a ZF filter and the negative-to-positive zero-crossings of the filtered signal are hypothesized as an estimate of GCIs.

The procedure to fit the LF model to the glottal source, $\dot{g}(t)$, using an exhaustive search method and dynamic programming is done using the voice analysis toolkit [57] as given in the next sub-section. The accuracy of the GCI estimation algorithm will affect the $F_0$ estimate and the time period for which the LF-model is fitted to the glottal estimate. However, the search method for the $R_d$ parameter and dynamic programming implementation uses estimated GCI locations from any given algorithm and then corrects the GCIs to match the main excitations of the $\dot{g}(t)$ (i.e., at its most negative peak in each glottal cycle). Thus, the GCI locations are aligned and adjusted to coincide with glottal source excitation minima. Hence, the dependency on the GCI extraction algorithm will be reduced. As an alternative to estimating the GCI locations from the speech signal, the $\dot{g}(t)$ estimated using the IAIF method can also be used directly for GCI estimation. However, this approach will require thresholding and peak picking of the negative peaks in the estimated $\dot{g}(t)$.

### *2) The exhaustive $R_d$ search algorithm*

In [50], a search algorithm was proposed in which the LF-model is estimated for all possible $R_d$ and the best $R_d$ is searched that minimizes the cost in the time and frequency domains. In the estimate of frequency-domain error, $H_g$ and $H_c$ are the harmonic spectra of $\dot{g}(t)$ and $g_c(t)$, respectively. The harmonic amplitudes are measured up to a frequency of *3 kHz*. The weights $w_t$, $w_s$ and $w_{tr}$ are associated with the time-domain error, frequency-domain error and transition cost, respectively. The steps to determine $R_d$ are given in Algorithm 1 and its MATLAB implementation is available in [57]. Normally, the $R_d$ value falls in the range $0.3 < R_d < 2.7$ whereas the upper range, i.e., $2.7 < R_d < 5$ signifies abduction. The $R_d$ parameter is known to govern all the other $R$-parameters [33]. Thus, from the $R_d$ parameter, the $R$-parameters are obtained as [33]:

$$R_a = (\text{-}1 + 4.8R_d)/100, \qquad R_k = (22.4 + 11.8R_d)/100,$$
$$R_g = 1/(4\times((0.11R_d/(1/2+1.2R_k))\text{ - }R_a)/R_k). \tag{7}$$

**Algorithm 1**: Exhaustive search algorithm to estimate $R_d$. After [50].

```
Step 1    For each GCI centered frame ⇒ $R_d = 0.3 : 0.1 : 5$
Step 1a   Use $F_0$ and $E_e$ for each $R_d$
          Time-domain error $Tr = \{0.5 - |corr\{\dot{g}(t), g_c(t)\}|\}.w_t$.
Step 1b   Frequency-domain error $Sr = \{0.5 - |corr\{Hg, Hc\}|\}.w_s$,
Step 1c   Total Error $Tot\_err = Tr + Sr$.
Step 2    Choose the five best candidates ($N_{cand}$) that minimize $Tot\_err$,
Step 3    The transition cost is,
          $\delta_{i,j,k} = \{0.5\text{-}|corr\{seg_{i,j}, seg_{i\text{-}1,k}\}|\}.w_{tr}$,
          $1 < j < N_{cand}$
          $1 \le i \le M$, where $M$= GCIs or analysis frames
Step 4    Optimal $R_d$ → minimize $D_{i,j} = d_{i,j} + \min\{D_{i-1,k}, \delta_{i,j,k}\}$.
```

**corr* is the correlation between the given variables.

Using the $R$-parameters, the $OQ$ can be estimated from eq. (5). Thus, we consider five shape features, i.e., $R_d$, $R_g$, $R_k$, $R_a$ and $OQ$. Fig. 3 shows the variations of the $R_d$ parameter for a male *Speaker A* and female *Speaker B*. For both the speakers, all the *150* natural speech utterances and *100* utterances each for *S1* VC and *S3* SS spoof from the training set of ASV spoof 2015 challenge database are used. It is observed from Fig. 3 that, for a speaker, the $R_d$ variations are different for natural, VC and SS spoof. Thus, speaker-specific properties are not exactly preserved when speech is synthesized or converted. For example, much higher values of $R_d$ are observed for *Speaker B* for SS than for natural and VC speech, indicating that SS sounds more breathy. In addition, across the speakers, the $R_d$ variations were different, signifying that $R_d$ captured speaker-related information both across natural and spoofed speech.

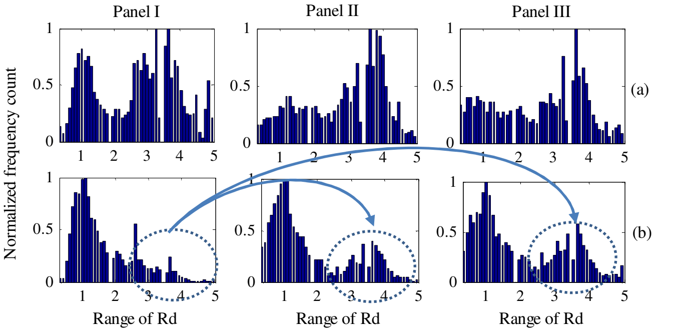

Fig. 3. Normalized histograms of the $R_d$ parameter for natural speech (Panel I), VC speech (Panel II) and SS (Panel III) corresponding to (a) *Speaker A* and (b) *Speaker B*. Dotted regions indicate the rise in $R_d$ value for VC and SS.

## IV. Features Based on Residual Information

This section describes the procedure to estimate the residual, $g_r(t)$, followed by the parameterization of the residual used for the SSD task.

### A. *Residual in the time domain*

Once the coarse structure $g_c(t)$ is fitted to the estimated $\dot{g}(t)$, the residual waveform is obtained as [29]:

$$g_r(t) = \dot{g}(t) - g_c(t). \tag{8}$$

The residue obtained from $\dot{g}(t)$ and $g_c(t)$ can be divided into *ripple* and *aspiration* components [29]. The ripple is known to have a frequency close to that of the first formant ($F_1$) of the vocal tract. It has also been shown that the ripple structure carries the speaker-specific information and, hence, is possibly a reason for the improvement in performance of speaker identification systems [44], [49]. On the other hand, aspiration occurs due to turbulence created at the glottis when airflow passes through the partially open glottis. The amount of aspiration contributes to the quality of voice (e.g., breathy). Using the $R$-parameters estimated from eq. (7), the values of the timing parameters ($t_o$, $t_e$ and $t_c$) are obtained. Considering the closed phase [0, $t_o$] for first glottal cycle or [$t_{c-1}$, $t_o$] for remaining glottal cycles, open phase [$t_o$, $t_e$] and the return phase [$t_e$, $t_c$], the energy (i.e., $L^2$ norm) of the residual $g_r(t)$ corresponding to these regions is denoted as $E_1$, $E_2$ and $E_3$, respectively. For each of the glottal cycles, the energy measurements are averaged over the glottal cycle, i.e.,

$$E_1 = \frac{\widehat{E}_1}{E_{tot}}, \quad E_2 = \frac{\widehat{E}_2}{E_{tot}}, \quad E_3 = \frac{\widehat{E}_3}{E_{tot}}, \tag{9}$$

where $\widehat{E}_1 = \int_0^{t_o} |g_r(t)|^2 dt$ or $\int_{t_{c-1}}^{t_o} |g_r(t)|^2\, dt$, $\widehat{E}_2 = \int_{t_o}^{t_e} |g_r(t)|^2\, dt$,

$\widehat{E}_3 = \int_{t_e}^{t_c} |g_r(t)|^2\, dt$ and $E_{tot} = \int_{t=0}^{T_0} |\dot{g}(t)|^2 dt$ is the total energy of $\dot{g}(t)$ in a glottal cycle. For a glottal cycle, Fig. 4 shows its corresponding estimated $\dot{g}(t)$, fitted LF-model $g_c(t)$ and its residual energy in closed phase, open phase and return phase,

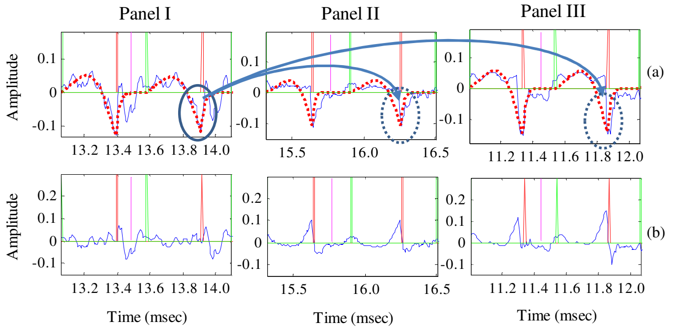

Fig. 4. For a voiced region of speech corresponding to natural speech (Panel I), VC (Panel II) and SS (Panel III), (a) estimated $\dot{g}(t)$ and its corresponding fitted LF-model, $g_c(t)$ and (b) ripple in the time domain, $g_r(t)$. The glottal opening (green), GCI location (red) and glottal closing location (magenta) indicate corresponding intervals for energies $E_1$, $E_2$ and $E_3$, respectively. The continuous oval in Panel I (a) indicates a close match between $\dot{g}(t)$ and $g_c(t)$, whereas the dotted region in Panel II (a) and III (a) indicates more deviation in the fit of $g_c(t)$ to the estimated $\dot{g}(t)$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2682788, IEEE Journal of Selected Topics in Signal Processing

corresponding to $E_1$, $E_2$ and $E_3$, respectively. It can be observed from Fig. 4 that during the open phase, the $\dot{g}(t)$ does not close gradually for spoofed speech as compared to natural speech due to which the residue in the open phase is large for the spoofed speech (as shown by the arrows in Fig. 4). In the regions other than open phase, the aspiration component is less for spoofed speech than in the natural speech signal.

As the ripple component exists in the open phase with corresponding energy $E_2$, we show the variations of the $E_2$ for *Speaker A* and *Speaker B* for all *150* natural speech utterances and all *100* utterances each for *S1* VC and *S3* SS spoof from the training set. It is observed from Fig. 5 that ripple energy in the open phase increases for VC as well as SS speech and across the speakers as well. Thus, the use of ripple energy has significant contribution in the SSD task.

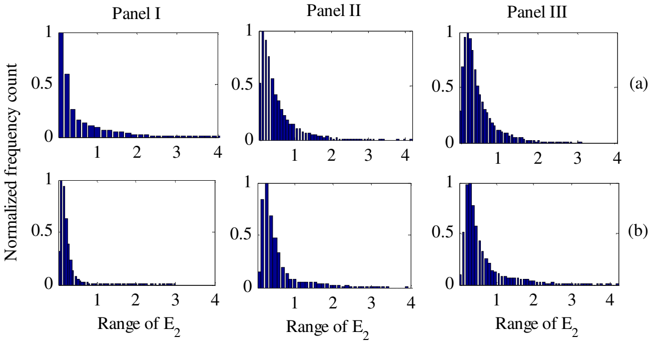

Fig. 5. Normalized histograms of $E_2$ for natural speech (Panel I), VC (Panel II) and SS (Panel III) corresponding to (a) *Speaker A* and (b) *Speaker B*.

### 1) *Variation of shape and energy features across speakers*

Next, to analyze the speaker-dependency of the five shape parameters ($R_d$, $R_g$, $R_k$, $R_a$ and $OQ$) and the three energy features ($E_1$, $E_2$ and $E_3$), we consider the entire training set of the ASV spoof 2015 dataset [10]. For the *25* speakers of the training data, the mean and standard deviation of the shape and energy parameters across *150* utterances of human speech and *100* utterances for five spoofs are considered here. We consider three shape parameters, $R_d$, $R_g$, and $OQ$ (shown in the top panel of Fig. 6) and three energy features, $E_1$, $E_2$ and $E_3$ (shown in the bottom panel of Fig. 6). The two shape parameters, $R_k$, $R_a$ are linearly related to $R_d$ and hence, not shown here. The blue 'o' and red '*' represent the points corresponding to the mean and standard deviations of natural and spoofed speech, respectively.

Firstly, considering the shape features, the variation across the speakers is more in natural speech than the spoofed speech for $R_d$ and $OQ$ parameters than for the $R_g$ parameter. $R_d$ for human speakers has spread across the normal range, $1 < R_d < 3$. However, the range for the spoofed speech was generally towards higher values of $R_d$, indicating more breathy voice for spoofed speech (generally, vocoder-based speech). Not much inference could be drawn from the $R_g$ parameter as the clusters of natural and spoofed speech were highly overlapping. On the other hand, the $OQ$ parameter goes in line with $R_d$, i.e., the $OQ$ had a higher mean and higher standard deviation across all the speakers for spoofed speech as compared to human speech. Secondly, the energy features were found to represent more distinctive features for natural and spoofed speech compared to the shape features. The mean of energy $E_1$ in the closed phase is less for spoofed speech than the natural speech. In the case of natural speech, this energy is referred to as *aspiration*, which is generally noise-like (due to creation of turbulence at the vocal folds). Next, as per the observations in Fig. 4 and Fig. 5, for $E_2$, the mean energy in the open phase is also more for spoofed speech than the natural speech. The energy $E_3$ in the return phase was less varied for natural speech than for the spoofed speech. Thus, the nature of variation in shape and energy features may result in features that may help to discriminate between the natural and spoofed speech.

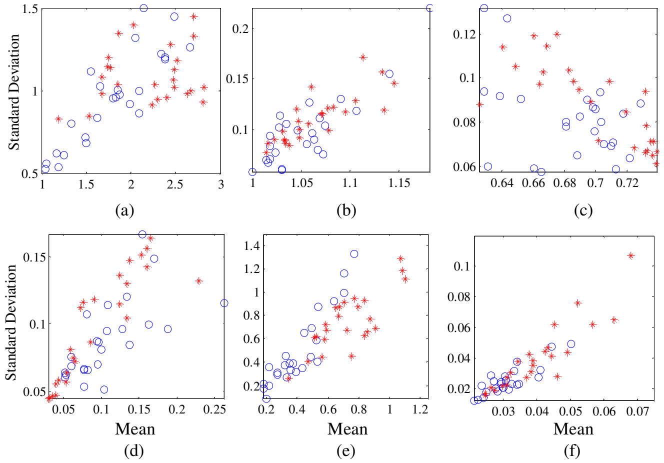

Fig. 6. The variations in terms of mean and standard deviation. Top Panel: for three shape parameters (a) $R_d$, (b) $R_g$, and (c) $OQ$ and Bottom Panel: for three energy features (d) $E_1$, (e) $E_2$ and (f) $E_3$ across all the speakers of the training dataset. Blue 'o' corresponds to speakers of the natural speech and red '*' corresponds to the spoofed speech for the same speakers.

## B. *Mel representation of the residual in the time domain*

The ripple component of $\dot{g}(t)$ can be due to the interaction of the excitation source with first formant ($F_1$), which is generally within *0-1 kHz* range. Thus, the information about the ripple is mainly embedded in the low frequency regions. Therefore, as an additional measure to enhance the ripple information in the residual signal, $g_r(t)$, we consider using the Mel cepstral representation of $g_r(t)$ as features for the SSD task. In addition, it was observed that the residual, $g_r(t)$, is intelligible and, hence, the use of Mel scale that more closely mimics the human perception process for hearing (than linearly-spaced frequency bands) can be used for estimating the subband energy of this excitation source signal. Such representations of using the Mel cepstra for the excitation source, such as the $\dot{g}(t)$ [44] and the LP residual [58], has been used for speaker identification task as well. Fig. 7 shows a speech signal, residual $g_r(t)$ and the Mel cepstral representation $g_r(t)$ for natural, VC speech and SS. The VC and SS correspond to the *S1* and *S3* of the ASV spoof 2015 database, respectively. As observed for natural speech in Fig. 7 (Panel I (c)), both low and high frequency regions are of high intensity as compared to spoofed speech. Therefore, the ripple component and the information about the aspiration component are more prominent in the natural speech as compared to the spoofed speech shown in Fig. 7.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2682788, IEEE Journal of Selected Topics in Signal Processing

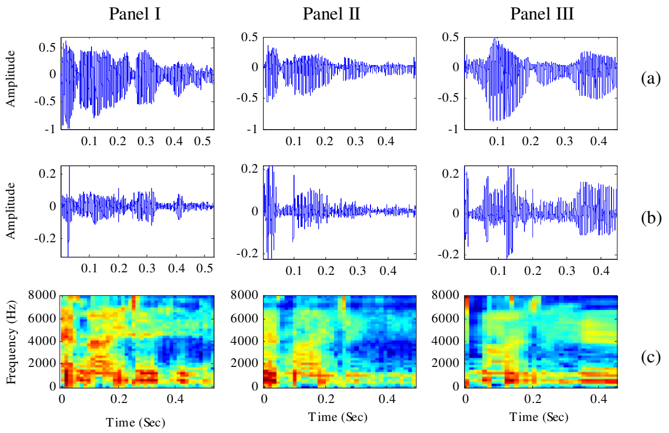

Fig. 7. (a) Speech signal (b) residual estimated from the difference of $\dot{g}(t)$ and $g_c(t)$ and (c) the Mel representation of (b) for natural speech (Panel I), VC speech (Panel II) and SS (Panel III).

### C. *Residual in the frequency domain*

To estimate the time-domain representation of the effect of S-F interaction, Ananthapadmanabha and Fant considered the vocal tract system as fixed and mapped all the nonlinear S-F interaction onto the excitation source [43], [44], [48]. The ripple in the time domain as a result of S-F interaction can be approximated by the following expression of $\dot{g}(t)$ [48],

$$\dot{g}(t) \approx g_c(t) + f(t)e^{-0.5tB_1(t)}\cos[\int_0^t F_1(\tau)d\tau], \qquad (10)$$

with $f(t)$ as the amplitude modulation (AM) and its multiplier reveals the first formant modulation (via *-3* dB bandwidth $B_1$ and the first formant ($F_1$) frequency) in the frequency domain. There exists a *duality* of ripple in the time domain and formant modulation in the frequency domain [43]- [48]. Thus, similar to the residue in the time domain, the residue in the frequency domain has significant information of the nonlinearities due to the S-F interaction. The concept of the residue in the frequency domain, (i.e., $Fr(\omega)$) is as follows,

$$\begin{aligned} Fr(\omega) &= 10\times\log\left|F\{\dot{g}(t)\}\right|^2 - 10\times\log\left|F\{g_c(t)\}\right|^2, \\ &= 10\times\log\left|\frac{F\{\dot{g}(t)\}}{F\{g_c(t)\}}\right|^2, \end{aligned} \qquad (11)$$

where $F\{\dot{g}(t)\}$ and $F\{g_c(t)\}$ is the Fourier transform of the $\dot{g}(t)$ and $g_c(t)$, respectively. Even though we intend to compute the residual-like representation of eq. (8) in the frequency domain, from eq. (11) it is clear that $Fr(\omega)$ represents the power ratio of the spectrum of $\dot{g}(t)$ and $g_c(t)$. Thus, the residue in the frequency domain will have formant modulation information at a much higher dimension than in the time domain.

Panel I and Panel II of Fig. 8 show the spectrograms of the estimated $\dot{g}(t)$ and fitted LF-model $g_c(t)$. The difference between these two spectra (as per eq. (11)) is shown in Fig. 8 (Panel III). It is observed that the energy spreads for natural speech (Fig. 8a) and spoofed speech (Fig. 8b and Fig. 8c) in Panel III are different for the low and high frequency regions. As in eq. (11), the residue of the spectrogram is also the power ratio of the spectrum of $\dot{g}(t)$ and $g_c(t)$. In the natural speech signal, as the intensity of the spectrum of $g_c(t)$ is high across the entire spectrum, the residual energy is least across all the frequency regions as compared to spoofed speech (Panel III). To further enhance the energy variations along the frequency-axis and to use it as features, the residual spectrogram (Panel III) is divided into *36* equally spaced regions and the energy is averaged over these regions. The representation is shown in Fig. 8 (Panel IV). This block-based energy will be further used for classification purposes in order to consider the effect of low and high frequency regions for the SSD task. Next, considering the fact that the ripple effect is mainly towards the lower frequency region, a better approach to enhance the lower frequency regions would be to use the Mel cepstral representation of the estimated $\dot{g}(t)$, the Mel cepstral representation of the fitted LF-model $g_c(t)$ and then obtain the residual in the frequency domain.

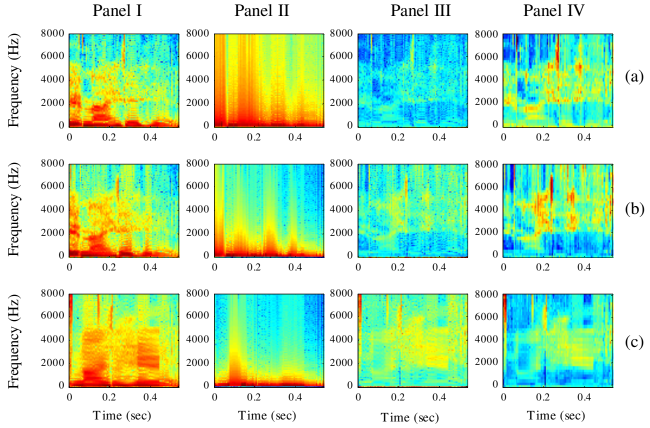

Fig. 8. The spectrogram of $\dot{g}(t)$ (Panel I), the spectrogram of the LF-model $g_c(t)$ (Panel II), the residue in the frequency domain, i.e., difference between spectrograms of $\dot{g}(t)$ and $g_c(t)$ (Panel III) and block-based energy of residual in the frequency domain (Panel IV) for (a) natural speech, (b) VC and (c) SS.

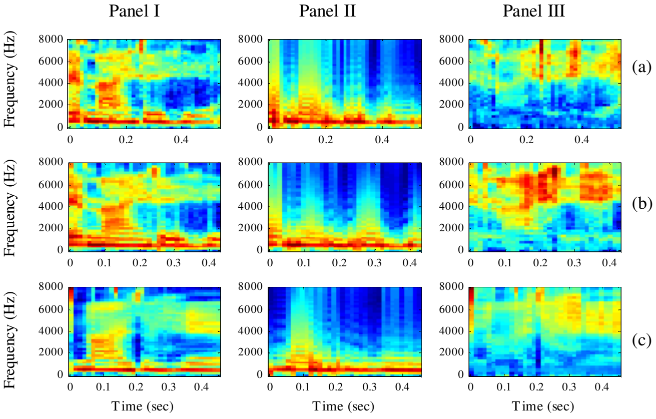

Fig. 9. The Mel representation of $\dot{g}(t)$ (Panel I), Mel representation of the LF-model $g_c(t)$ (Panel II), the residue in the frequency domain, i.e., difference between Mel representations of $\dot{g}(t)$ and $g_c(t)$ (Panel III) for the (a) natural speech, (b) VC speech and (c) SS.

Fig. 9 shows the Mel representation of the estimated $\dot{g}(t)$ and the fitted LF-model $g_c(t)$ in Panel I and Panel II, respectively. The difference between the two representations (shown in Fig. 9 (Panel III)) is considered as the residue of Mel-warped representations. We observe that the difference in the residue of Mel representations of $\dot{g}(t)$ and $g_c(t)$ for natural and spoofed speech differ across the utterances. In both the lower and higher frequency region, the energy for the natural speech was less than that of the spoofed speech. The closely spaced filters in the lower frequency region enhances features essential for spoof detection, which on fitting well for natural speech results in less energy in lower frequency regions.

## V. Experimental Setup and Results

In this section, the databases and the performance measures are discussed. The experimental setup is the same as in our companion paper in the same special issue [59]. Next, the results on the development set are shown, followed by results on the evaluation set (with unknown spoofing algorithms) for the clean case, noisy and channel mismatch conditions.

### *A. Spoofed Speech Detection (SSD) System Setup*

The ASVspoof 2015 challenge database consists of training and development (with *S1–S5* as spoofing algorithms) and evaluation set (with *S1–S10* as spoofing algorithms). The *S1-S9* are vocoder-dependent while *S10* vocoder-independent. Other technical details about the spoofing technique and type of vocoder used are given in [10]. Next, to evaluate the performance of the features in channel mismatch scenarios, the Blizzard Challenge 2012 database is considered [54]. There are *11* systems, from *A* to *K* with each system having *100* sentences. In particular, system A contains natural speech signals whereas systems *B*, *G*, *F* and *I* were built using the USS approach. Systems *E*, *H*, *K* were built using statistical methods. Systems *C* and *D* were built using a hybrid approach and *J* was built using the diphone-based method.

A Gaussian Mixture Model (GMM)-based classifier with *128* mixtures models is used. Final scores on a test sequence *Y* in terms of a log-likelihood ratio (*LLR*) are represented as,

$$LLR = \log\ (p\ (Y\,|\,\lambda_{nat})) - \log\ (p\ (Y\,|\,\lambda_{syn})), \qquad (12)$$

where $p(Y|\lambda_{\text{nat}})$ and $p(Y|\lambda_{\text{syn}})$ are the likelihood scores from the $\lambda_{\text{nat}}$ (GMM for natural speech) and $\lambda_{\text{syn}}$ (GMM for spoofed speech), respectively. To utilize possible complementary information, score-level fusion of features is done, i.e.,

$$LLk_{fused} = (1-\alpha_f)LLk_{feature1} + \alpha_f LLk_{feature2}, \qquad (13)$$

where $LLk_{fused}$ is the fused log-likelihood score of *feature1* and *feature2*. We also consider fusing the source features (both in time and frequency domains) and the system features all together using the following equation,

$$LLk_{fused} = \alpha_1 LLk_{feature1} + \alpha_2 LLk_{feature2} + \alpha_3 LLk_{feature3}, \qquad (14)$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$ and *LLk* scores of *feature1*, *feature2 and feature 3* are fused at the score-level. The score-level fusion gives the contribution of the individual features and avoids higher feature dimension due to feature-level fusion.

Finally, the Detection Error Tradeoff (DET) curve is used to study the SSD system performance. The operating point in the DET curve where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal is referred to as the Equal Error Rate (EER) [60]. The average % EER is estimated by considering the natural speech as the positive class and all spoofed speech from all the spoofing algorithms as the negative class [59]. Interestingly, this attack-independent approach to estimate % EER is also used in the upcoming ASVspoof 2017 challenge [61].

### *B. Parameterization*

To derive the features out of the LF-model, we use the five shape-related features ($R_d$, $R_g$, $R_k$, $R_a$ and *OQ*) and the three energy features ($E_1$, $E_2$ and $E_3$) in the time domain. In the frequency domain, we use a Mel representation of the residual, i.e., $g_r(t)$ (as defined in eq. (8)). In addition, the residual of the spectrogram of the estimated $\dot{g}(t)$ and the fitted LF-model $g_c(t)$ also effectively represent the features for the SSD task. Here, a *256*-point FFT is considered. As shown earlier, the frequency-axis is divided into *36 equally* spaced regions and the energy is averaged over these regions. Thereafter, we explore whether the Mel representation (instead of only FFT) provides better features for the SSD task. In this context, the Frequency-domain residual Feature Representation (*FrFR*) is summarized in Table I. The Discrete Cosine Transform (DCT) of the representations is taken to obtain static (without the $0^{th}$ energy coefficient) and dynamic features, (i.e., delta (Δ) and delta-delta (ΔΔ)). Thus, feature vectors, *D1*:*12-D* static features, *D2*: *24-D* (*12*s+*12*-Δ), *D3*:*36-D* (*12*s+*12*-Δ+*12*-ΔΔ) are considered.

TABLE I
The Frequency-domain Residual Feature Representations (FrFR)

| *FrFR* | **Feature Sets** | **Dimensions (D)** |
|---|---|---|
| *FrFR1* | The static and dynamic representations of the Mel cepstral representation of the residue $g_r(t)$ | *12*s+*12*Δ+*12*ΔΔ |
| *FrFR2* | The block-based energy of residue of spectrogram of $\dot{g}(t)$ and $g_c(t)$ | *36-D* |
| *FrFR3* | The static and dynamic representations of *FrFR2* | *12*s+*12*Δ+*12*ΔΔ |
| *FrFR4* | The equally divided low frequency region (LFR) and high frequency regions (HFR) of *FrFR2* | LFR: *18-D*<br>HFR: *18-D* |
| *FrFR5* | The static and dynamic representations of residue of Mel cepstra of $\dot{g}(t)$ and $g_c(t)$ | *12*s+*12*Δ+*12*ΔΔ |

*s=static, Δ=delta, ΔΔ=delta-delta.*

For extracting the GCI locations, as discussed, the ZF method is used. A frame size of *30 ms* and a frame shift of *10 ms* are considered. To estimate the LF-model from the $R_d$ search algorithm, equal weights $w_t$, $w_s$ and $w_{tr}$ are associated with the time-domain error, frequency-domain error and transition cost, respectively. From the fitted LF-model, it was observed that for the weak voiced regions, an epoch may not always be estimated, due to which the LF-model may be ill-fitted. This resulted in the outliers in the estimated energy values in the closed, open and return phase which needs to be discarded during the training of GMMs. Thus, for time-domain features, extreme data points outside the $1^{st}$ percentile to the $99^{th}$ percentile are discarded. This is also done to alleviate the possible components of GMM that might model an outlier distribution (especially in the case of use of a large number of mixtures in GMM). Moreover, the presence of such outliers in features may shift the mean and the variance of component

TABLE II
% EER ON THE DEVELOPMENT SET OF VARIOUS FREQUENCY-DOMAIN RESIDUAL FEATURE REPRESENTATION (FRFR) FOR THE SSD TASK

| Feature Set | FrFR1 | | | FrFR2 | FrFR3 | | | FrFR4 | | | FrFR5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dimension of FV | D1 | D2 | D3 | 36-D | D1 | D2 | D3 | 18-D<br>(LFR) | 18-D<br>(HFR) | Score-level Fusion<br>(LFR+HFR) | D1 | D2 | D3 |
| **% EER** | 10.06 | 8.17 | **7.80** | 23.65 | 26.85 | 18.93 | 15.27 | 24.02 | 28.96 | 21.90 | 12.30 | 9.67 | **8.81** |

*FV= Feature Vector,* LFR = low frequency region (*0-4 kHz*)→feature 1, HFR = high frequency region (*4-8 kHz*)→feature 2, *D1=12s, D2=12s+12Δ, D3=12s+12Δ+12ΔΔ.*

TABLE III
% EER ON THE DEVELOPMENT SET FOR SCORE-LEVEL FUSION OF *FRFR1, FRFR5*, ENERGY AND SYSTEM FEATURES AT VARIOUS FUSION FACTORS AS IN EQ. (13)

| Feature Set 1 | Dimension of FV | Fusion Factor ($\alpha_f$) | | | | | | | | | | | Dimension of FV | Feature Set 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1.0** | | |
| **FrFR1** | D1 | 10.06 | 9.29 | 8.52 | 8.09 | 7.69 | **7.55** | 7.75 | 8.09 | 8.89 | 10.41 | 12.30 | D1 | **FrFR5** |
| | D2 | 8.17 | 7.43 | 6.83 | 6.43 | 6.26 | **6.23** | 6.32 | 6.66 | 7.35 | 8.29 | 9.67 | D2 | |
| | D3 | 7.80 | 7.12 | 6.52 | 6.06 | 5.80 | **5.75** | 5.89 | 6.23 | 6.81 | 7.75 | 8.81 | D3 | |
| **Energy** | 3-D | 6.09 | 3.97 | **3.46** | 3.60 | 4.12 | 4.78 | 5.40 | 6.09 | 6.78 | 7.26 | 7.81 | D3 | **FrFR1** |
| **Energy** | 3-D | 6.09 | 4.29 | **3.80** | 3.92 | 4.35 | 5.03 | 5.80 | 6.66 | 7.41 | 8.15 | 8.81 | D3 | **FrFR5** |
| **Energy** | 3-D | 6.09 | 1.77 | 0.80 | 0.46 | **0.43** | 0.51 | 0.66 | 0.83 | 1.14 | 1.34 | 1.60 | D3 | **MFCC** |
| **FrFR1** | D3 | 7.81 | 4.89 | 3.06 | 1.74 | 1.09 | 0.83 | **0.69** | 0.80 | 0.92 | 1.32 | 1.60 | D3 | **MFCC** |
| **FrFR5** | D3 | 8.81 | 5.66 | 3.37 | 2.06 | 1.29 | 0.92 | **0.74** | 0.89 | 1.03 | 1.34 | 1.60 | D3 | **MFCC** |

*FV= Feature Vector, D1=12s, D2=12s+12Δ, D3=12s+12Δ+12ΔΔ.*

Gaussians used in GMM. In addition, the presence of outliers in training or testing results in mis-classification, thereby resulting in an increase in EER.

### *C. Results on the Development Set*

Next, as the time-domain features of shape and energy are less in dimension, we experiment to find the relatively optimal number of Gaussian mixture components that would be required to model the features. Therefore, as shown in Fig. 10, we train the GMM on various numbers of mixture components and test it using the development set.

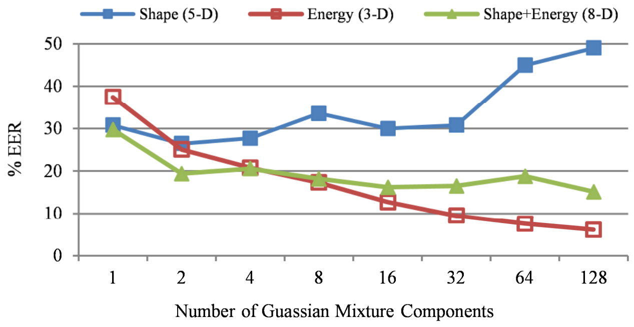

Fig. 10. The % EER for *5-D* shape features, *3-D* energy features and the *8-D* combination of shape and the energy features at feature-level for a various number of Gaussian mixture components varied from *1* to *128*.

It is observed that for the shape features, no improvement in the performance was observed with the increase in the number of mixture components. However, for energy features, the % EER decreased significantly with the mixture components. Even on using the shape and energy features together at feature-level, the % EER of the energy-based features alone was less. A possible reason for shape features not performing well is that the $R_d$ parameter is limited to the range of *0.3* to *5* and other *R*-parameters are derived from $R_d$ itself. On the other hand, the energy features result due to differences between estimated and fitted model signifying the nonlinear S-F interaction. Thus, there is more possibility of capturing the differences between natural and spoofed speech. Hence, for clean speech, we consider the use of only $E_1$, $E_2$ and $E_3$ energy values as the time-domain features using a GMM of *128* mixture components. Table II shows the % EER for the *FrFR* features. Few of the observations can be summarized as follows:

- For *FrFR1*, the EER is *10.06%* for the *D1* feature vector and reduces to as low as *7.80%* using the *D3* feature vector. Thus, dynamic variations in feature trajectories are found to be effective for the SSD task.
- Next, using *FrFR2*, i.e., the *36-D* block-based energy representation yields an EER of *23.65%*.
- Use of DCT of *FrFR2* to obtain *FrFR3*, reduces the EER from *26.85%* for *D1* feature vector to *18.9%* and *15.27%* for *D2* and *D3* feature vectors, respectively.
- In order to investigate which frequency regions capture spoof-specific characteristics, the *FrFR2*, i.e., *36-D* block-based energy representation as in Fig. 8 (Panel IV) is divided into two frequency regions, called *FrFR4*, i.e., *18-D* low frequency region (LFR) for the range *0-4 kHz* and *18-D* high frequency region (HFR) for *4-8 kHz*. The EER of LFR as a feature set is found to be *24.02%* which is less than *28.96%* when the HFR are used and less than *12-D FrFR3* static features. On score-level fusion of LFR and HFR (with $\alpha_f$*=0.3* as in eq. (13)), the EER obtained is *21.9%*, which is less than that obtained by LFR. This implies that it may be appropriate to process the residue of the spectrograms such that the LFR are enhanced more than the HFR.
- To enhance the LFR, we consider *FrFR*5, i.e., the residue of the Mel representation of the estimated $\dot{g}(t)$ and fitted LF-model $g_c(t)$. The EER obtained with this is *12.30%* for *D1* feature vector, which reduces to *9.67%* and *8.81%* with the addition of Δ and ΔΔ features.

Hence, for the remaining set of experiments, we consider the *FrFR1* and *FrFR5* as feature vectors for the SSD task. These EERs are not less than *6.09%*, which is achieved for *3-D* time-domain energy-based features.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2682788, IEEE Journal of Selected Topics in Signal Processing

### 1) *Score-level fusion of source features*

To consider *jointly* the effect of any two feature sets, we perform a score-level fusion of the features. The fusion of two features and for three features is done as in eq. (13) and in eq. (14), respectively. It is observed in Table III that when the *FrFR1* is fused with *FrFR5*, the least EER obtained is *5.75%* using *D3* feature vector. This EER is less than *FrFR1* and *FrFR5* used alone and also less than *6.09%* of the *3-D* energy features in the time domain. The energy-based features when fused with the *D3* feature vector of the *FrFR1* and *FrFR5* (with $\alpha_f$=*0.*2), a significant improvement in performance is obtained resulting in *3.46%* and *3.80%* EER, respectively. The DET curves for the energy features, *FrFR1* features, *FrFR5* features and their score-level are shown in Fig. 11.

It is observed that without fusion, the *FrFR5* has the highest % EER and the energy-based features have the least % EER. However, the FAR of the time-domain energy features were far more than *FrFR1* and *FrFR5* for FRR less than *2%*. Both the *FrFR1* and *FrFR5* features had more % FRR than the time-domain energy-based features. Thus, it is a feasible option to fuse the time-domain energy-based features and frequency-domain representations for better performance. The score-level fusion of time-domain energy features with *FrFR1* and *FrFR5* features decreased the % EER as well as the % FRR significantly. It was observed that after fusion, the % FAR did not reduce much for less than *0.5%* FRR.

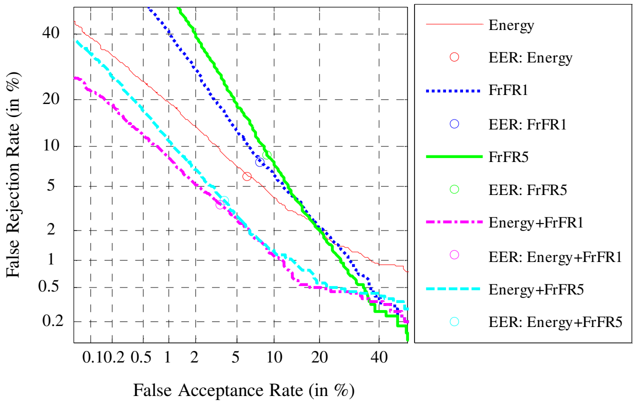

Fig. 11. The DET curve on the development set for *3-D* time-domain energy features (red), *FrFR1* (blue), *FrFR5* (green), score-level fusion ($\alpha_f$=*0.2*) of *3-D* energy features with *FrFR1* (magenta) and *FrFR5* (cyan).

### 2) *Score-level fusion of source and system features*

Next, we consider the system-based MFCC feature set. The MFCC features were extracted using *28* subband Mel filters, with a frame size of *25 ms* and with *50%* overlap [11]. On the development set, the MFCC features with *D3* feature vector gave an EER of *1.6%*. With MFCC as the system-based feature set, we fuse at the score-level the source-based information, namely, energy features, *FrFR1* and *FrFR5*. Therefore, as shown in Table III, upon fusing the energy features with MFCC at $\alpha_f$ = *0.4,* the EER drops down to as low as *0.43%*. Fusing *FrFR1* and *FrFR5* at the score-level with MFCC at $\alpha_f$ = *0.6*, the EER decreases to *0.69%* and *0.74%*, respectively. Thus, more complementary information is clearly present in the source-based features than MFCC alone for the SSD task. Moreover, the greater weightage to energy-based features indicates their relative importance than system-based features. Next, to use the system features and source features (time-domain and frequency-domain features), an equal weight factor of $\alpha_f$ *=0.5* is used. Therefore, for factors $\alpha_1$*=0.4*, $\alpha_2$*=0.1* and $\alpha_3$*=0.5* (as in eq. (14)), EERs of as low as *0.371%* and *0.457%* are obtained as shown in Table IV. This % EER is much less than the best EER of *0.83%* submitted at the ASV spoof 2015 challenge using a score-level fusion of CFCCIF and MFCC features. Hence, it is clear from the experiments that the S-F interaction residual energy features in the time domain and frequency domain are highly essential to capture the vocoder-specific characteristics.

TABLE IV

%EER FOR SCORE-LEVEL FUSION OF SOURCE (TIME-DOMAIN ENERGY AND FREQUENCY-DOMAIN FEATURES) AND MFCC FEATURES AS IN EQ. (14)

| Score-level Fusion | % EER |
|---|---|
| $\alpha_1$ . Energy + $\alpha_2$ . FrFR1 + $\alpha_3$ . MFCC | 0.3717 |
| $\alpha_1$ . Energy + $\alpha_2$ . FrFR5 + $\alpha_3$ . MFCC | 0.4575 |
| $\alpha_1$*=0.4*, $\alpha_2$*=0.1* and $\alpha_3$*=0.5* | |
| EER submitted at the ASV spoof 2015 challenge, INTERSPEECH 2015 | 0.8293 |

### D. *Results of the Evaluation Set*

In the realistic scenarios, the type of spoof (SS or VC) or the spoofing algorithm will not be known. Thus, the performance is studied by testing with all *S1-S10* spoofs of the evaluation set. As analyzed from the development set, a score-level fusion of source and system-based features gave significantly lower % EER than the features used individually. Considering the features and the fusion factors obtained from the development set, the overall % EER on the evaluation set and the error of the individual *S1* to *S10* attacks are shown in Table V. Considering the EER for the known attacks, for the *3-D* time-domain energy features, the EER is *1.84%* which is significantly less. For the feature-level fusion (*8-D*) of the shape and energy-based features, the % EER is high and hence, not considered further. The *FrFR1* and *FrFR5* with *D3* feature vector gives *5.672%* and *6.805%* EER. A score-level fusion of energy features with *FrFR1* and *FrFR5* at $\alpha_f$ = *0.2* gives *0.736%* and *1.134%* EER, respectively. Secondly, considering the system-based MFCC features, the EER is found to be around *0.36%* for known attacks. On combining the energy and MFCC features at $\alpha_f$*=0.4*, the % EER reduced ten times compared to MFCC to achieve an EER of *0.034%*. Similar observations were found when *FrFR1* and *FrFR5* features were fused at score-level with MFCC. Thus, use of source-based features for SSD can be justified.

The evaluation set consists of both vocoder-dependent and vocoder-independent speech. Hence, the overall % EER is obtained without *S10* (only vocoder-based (*S1-S9*)) and with *S10*. The % EER of vocoder-based spoofs (*S1-S9*) is similar to that of the known case. That is, the vocoder-based attacks gave a significantly low EER of *0.017%* on the score-level fusion of the energy, *FrFR1* and MFCC features with factors of $\alpha_1$*=0.4*, $\alpha_2$*=0.1* and $\alpha_3$*=0.5*, respectively. An example of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2682788, IEEE Journal of Selected Topics in Signal Processing

TABLE V-A
% EER ON TESTING WITH THE EVALUATION SET FOR TIME-DOMAIN ENERGY FEATURES, *FRFR1* AND *FRFR5* FEATURES, SCORE-LEVEL FUSION OF TIME-DOMAIN ENERGY FEATURES, *FRFR1* AND *FRFR5* FEATURES AND MFCC AT FUSION FACTORS DECIDED FROM THE DEVELOPMENT SET

| **Feature Set** | **Dimension** | **Individual Spoofing Attacks (% EER)** | | | | | | | | | | **Overall % EER** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **S1** | **S2** | **S3** | **S4** | **S5** | **S6** | **S7** | **S8** | **S9** | **S10** | **Kn.** | **S1-S9** | **S1-10** |
| **Energy** | 3-D | 0.212 | 6.902 | 0.130 | 0.163 | 1.777 | 4.005 | 21.179 | 2.141 | 7.853 | 86.429 | 1.837 | 4.930 | 13.07 |
| **Shape + Energy** | 8-D | 4.174 | 17.05 | 9.277 | 9.120 | 14.59 | 21.76 | 26.908 | 3.342 | 10.65 | 81.087 | 10.84 | 12.99 | 19.79 |
| **FrFR1** | D1 | 7.033 | 14.543 | 1.832 | 1.897 | 8.462 | 13.120 | 5.152 | 1.348 | 3.946 | 84.304 | 6.752 | 6.370 | 14.16 |
| | D2 | 5.457 | 13.560 | 0.630 | 0.761 | 6.848 | 11.543 | 2.576 | 1.234 | 3.668 | 84.212 | 5.452 | 5.142 | 13.05 |
| | D3 | 5.772 | 15.212 | 0.842 | 0.761 | 5.783 | 10.174 | 2.098 | 1.755 | 3.853 | 82.750 | 5.672 | 5.137 | 12.90 |
| **FrFR5** | D1 | 4.136 | 21.853 | 6.348 | 6.848 | 8.745 | 10.082 | 27.451 | 12.473 | 7.397 | 77.120 | 9.586 | 11.70 | 18.25 |
| | D2 | 1.951 | 22.766 | 3.016 | 3.043 | 7.826 | 8.299 | 16.402 | 9.967 | 9.402 | 76.397 | 7.721 | 9.186 | 15.91 |
| | D3 | 0.962 | 21.359 | 2.283 | 2.337 | 7.087 | 7.842 | 8.375 | 8.560 | 7.353 | 76.761 | 6.805 | 7.351 | 14.29 |
| **Energy+FrFR1** | D3+3-D | 0.038 | 3.092 | 0.011 | 0.033 | 0.505 | 1.815 | 1.332 | 0.212 | 1.321 | 88.375 | 0.736 | 0.929 | 9.673 |
| **Energy+FrFR5** | D3+3-D | 0.054 | 4.815 | 0.043 | 0.065 | 0.690 | 1.511 | 5.272 | 0.826 | 2.457 | 86.761 | 1.134 | 1.748 | 10.25 |
| **MFCC** | D3 | 0.005 | 0.995 | 0.000 | 0.000 | 0.832 | 0.902 | 0.054 | 0.000 | 0.082 | 39.723 | 0.366 | 0.319 | 4.259 |
| **Energy+MFCC** | 3-D+D3 | 0.000 | 0.141 | 0.000 | 0.000 | 0.027 | 0.038 | 0.000 | 0.000 | 0.005 | 47.913 | 0.034 | 0.024 | 4.813 |
| **FrFR1+MFCC** | D3+D3 | 0.000 | 0.196 | 0.000 | 0.000 | 0.125 | 0.304 | 0.016 | 0.000 | 0.005 | 47.402 | 0.064 | 0.072 | 4.805 |
| **FrFR5+MFCC** | D3+D3 | 0.000 | 0.571 | 0.000 | 0.000 | 0.321 | 0.272 | 0.016 | 0.000 | 0.022 | 44.011 | 0.178 | 0.133 | 4.521 |
| **Energy+FrFR1+MFCC** | 3-D+D3+D3 | 0.000 | 0.065 | 0.000 | 0.000 | 0.027 | 0.060 | 0.000 | 0.000 | 0.000 | 42.261 | **0.018** | **0.017** | 4.241 |
| **Energy+FrFR5+MFCC** | 3-D+D3+D3 | 0.000 | 0.174 | 0.000 | 0.000 | 0.049 | 0.043 | 0.000 | 0.000 | 0.005 | 40.973 | 0.045 | 0.030 | **4.124** |

TABLE V-B
% EER ON TESTING WITH THE EVALUATION SET FOR COCHLEAR-BASED CFCC, CFCCIF, CFCCIFS FEATURES AND SOURCE-BASED FEATURES WHEN FUSED AT SCORE-LEVEL WITH MFCC FEATURE SET

| **Feature Set** | **Dimension** | **Individual Spoofing Attacks (% EER)** | | | | | | | | | | **Overall % EER** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **S1** | **S2** | **S3** | **S4** | **S5** | **S6** | **S7** | **S8** | **S9** | **S10** | **Kn.** | **S1-S9** | **S1-10** |
| **CFCC+MFCC** | D3+D3 | 0.016 | 0.740 | 0.000 | 0.000 | 1.330 | 0.680 | 0.076 | 0.000 | 0.120 | 13.080 | 0.417 | 0.329 | 1.604 |
| **CFCCIF+MFCC** | D3+D3 | 0.000 | 0.360 | 0.000 | 0.000 | 0.970 | 0.500 | 0.043 | 0.082 | 0.049 | 16.720 | 0.266 | 0.223 | 1.872 |
| **CFCCIFS+MFCC** | D3+D3 | 0.000 | 0.240 | 0.000 | 0.000 | 0.720 | 0.310 | 0.033 | 0.098 | 0.038 | 13.030 | 0.192 | 0.160 | 1.447 |
| **(F0-SoEs)+MFCC** | 12-D+D3 | 0.000 | 0.720 | 0.000 | 0.000 | 0.190 | 0.300 | 0.020 | 0.000 | 0.030 | 34.470 | 0.182 | 0.140 | 3.573 |
| **Best(M1+M2)+MFCC** | 24-D+D3 | 0.000 | 0.040 | 0.000 | 0.000 | 0.020 | 0.020 | 0.010 | 0.000 | 0.010 | 51.110 | 0.012 | 0.011 | 5.121 |

TABLE V-C
% EER ON TESTING WITH THE EVALUATION SET FOR SCORE-LEVEL FUSION OF ENERGY FEATURES, *FRFR1* AND *FRFR5* FEATURES WITH CFCCIFS FEATURE SET

| **Feature Set** | **FF** | **Individual Spoofing Attacks (% EER)** | | | | | | | | | | **Overall % EER** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **S1** | **S2** | **S3** | **S4** | **S5** | **S6** | **S7** | **S8** | **S9** | **S10** | **Kn.** | **S1-S9** | **S1-10** |
| **Energy+FrFR1+CFCCIFS** | A1 | 0.005 | 0.011 | 0.000 | 0.000 | 0.033 | 0.060 | 0.005 | 0.136 | 0.011 | 16.136 | 0.010 | 0.029 | 1.640 |
| **Energy+FrFR5+CFCCIFS** | A1 | 0.000 | 0.016 | 0.000 | 0.000 | 0.071 | 0.071 | 0.000 | 0.201 | 0.027 | 15.179 | 0.017 | 0.043 | 1.557 |
| **Energy+FrFR1+CFCCIFS** | A2 | 0.011 | 0.082 | 0.000 | 0.000 | 0.261 | 0.136 | 0.011 | 0.332 | 0.043 | 13.082 | 0.071 | 0.097 | 1.396 |
| **Energy+FrFR5+CFCCIFS** | A2 | 0.011 | 0.141 | 0.000 | 0.000 | 0.348 | 0.120 | 0.038 | 0.451 | 0.060 | **12.837** | 0.100 | 0.130 | **1.401** |

*Fusion Factors (FF) A1: $\alpha_1$=0.4, $\alpha_2$=0.1* and *$\alpha_3$=0.5, A2: $\alpha_1$=0.2, $\alpha_2$=0.1* and *$\alpha_3$=0.7*

effectiveness of score-level fusion is the *S7* spoof for which the time-domain energy features give an EER of *21.18%*. However, when fused at score-level with *FrFR1* and *FrFR5* features, the EER reduced to *1.332%* and *5.272%*, respectively. The EER for *S7* spoof decreases to *0.00%* on the score-level fusion of energy and MFCC features.

Considering the % EER of the *S10* spoof, it is observed that the relatively best % EER was obtained using the MFCC features. It is observed that the S-F interaction features do not contribute significantly when used alone or with the MFCC feature set. Slightly lower EER of *4.124%* than MFCC alone was obtained on fusing energy, *FrFR5* and MFCC features with fusion factors of *$\alpha_1$=0.4*, *$\alpha_2$=0.1* and *$\alpha_3$=0.5*, respectively. However, this decrement is not very significant. Thus, the S-F interaction features do not contribute in detecting the spoofed speech due to concatenative speech synthesis where, in principle, natural speech sound units are joined and hence, create a lot more confusion during classification of natural and spoofed speech. This is due to the fact that the residual features are characteristics of the natural speech that are still preserved in the *S10* spoof due to the direct concatenation of the natural speech sound units. Thus, the residual features in time and frequency domains may not prove to be very effective

### *E. Comparison with Existing Features*

The performance of the S-F interaction features is compared with the authors' previous work on using cochlear filter-based CFCC, CFCCIF and CFCCIFS feature sets as well as other excitation source features such as, *$F_0$* and *SoE* features and prediction residual-based features. These features were also fused at score-level with MFCC. A brief description of the features and their relative performance as compared to S-F interaction features is given as follows.

- *Cochlear-based features*: The CFCC feature sets are based on using the auditory filterbanks as compared to triangular filterbanks in MFCC. In addition, the envelope at the output of the cochlear subband is combined with the average subband IF information. Moreover, to capture transient information or the variation across the frames, the derivative operation is used. These features are known as CFCCIF [18]. In addition, the use of symmetric difference to estimate variations of subband energy representation (i.e., CFCCIFS) has shown to give better performance [59]. As shown in Table V-B, the cochlear-based features, when combined with MFCC using score-level (*$\alpha_f$=0.2* as per eq. (13)) gave a very low average % EER of *1.44%*. However, for the vocoder-based *S1-S9* spoofs, the EER was *0.16%* which is almost ten times more than with the energy and *FrFR1* features when fused with MFCC at score-level.

- *$F_0$ and SoE features:* These features are based on the fact that when the vocal folds vibrate, there exists a correlation between the $F_0$ contour and *SoE* at the glottal excitation source (*SoE1*) and at the speech signal (*SoE2*), which is found to be more for natural speech than machine-generated speech [18]. Moreover, as natural speech has more variations, the dynamics of the $F_0$, *SoE1* and *SoE2* features are also considered by taking their derivative up to $3^{rd}$ order. These features when combined with MFCC using score-level (*$\alpha_f$=0.8* as per eq. (13)) perform slightly better than cochlear-based features for *S1-S9* spoofs. However, even in this case for *S1-S9* spoofs, the % EER of proposed S-F features is almost ten times better.

- *LP-LTP and LP-NLP features*: Here, the LP, LTP and NLP features are explored based on the idea that the spoofed speech is too easy to predict if a simplified acoustic model generates it and it is too difficult to predict if there are artifacts present in the speech signal [19], [20]. Hence, the score-level fusion of LP-LTP (M1) and LP-NLP (M2) combination at *$\alpha_f$=0.4*, when further combined with MFCC at score-level (*$\alpha_f$=0.1* as per eq. (13)) provided discriminative or complementary features especially for *S1-S9* spoofs. The performance for only vocoder-based spoofs is slightly better than the S-F interaction features. However, the EER for prediction-based features is high for *S10*; as a result the average % EER is more than the proposed S-F interaction features.

- *Fusion of S-F interaction features with CFCCIFS feature set:* From Table V-A and V-B, it is observed that S-F interaction features work well in detecting the vocoder-based speech (*S1-S9*). On the other hand, the authors' previous work of using envelope and IF information jointly (i.e., CFCCIF and CFCCIFS features) gave reduced %EER for *S10* spoof as compared to S-F interaction or other source-based features. Therefore, we attempt to combine the benefits from both of these two features as shown in Table V-C. Firstly, for combination of time-domain energy features, FrFR1/FrFR5 and CFCCIFS features, the fusion factors are *$\alpha_1$=0.4*, *$\alpha_2$=0.1* and *$\alpha_3$=0.5*. It is observed that with the addition of CFCCIFS features, the EER of *S10* decreases to around *~15-16%*. On increasing the contribution of CFCCIFS features (i.e., *$\alpha_1$=0.2*, *$\alpha_2$=0.1* and *$\alpha_3$=0.7)*, the average EER decreases to around *~1.4 %* where the EER of *S10* decreases to *13.08* and *12.83* on using *FrFR1* and *FrFR5* features, respectively. The performance of *S1-S9* slightly degrades as compared to S-F interaction and MFCC features, however, the performance of *S1-S9* is better than CFCCIFS +MFCC system. It is to be noted that the use of attack-independent threshold makes the detection of *S10* even more difficult and hence responsible for large average EER.

### *F. Results on Signal Degradation Conditions*

Amongst the several approaches used for the SSD task, the results are mostly evaluated in the presence of clean conditions. Very recently, research had been directed towards evaluating the performance of countermeasures in the presence of noisy environments. In [62], a preliminary investigation of spoofing detection under additive noisy conditions had been performed. This work also describes an initial noisy database developed by artificially adding background noises at different SNR levels. The work shows that for a model trained on clean data, the system performance degrades significantly when tested on noisy speech. It was observed that the system performance differs with the types of noises. In [63], on similar grounds, several countermeasures were found to fail at relatively high SNRs and did not generalize well for the SSD task even with speech enhancement algorithms.

In this study, we consider evaluating the performance of the proposed S-F interaction features for additive white noise, babble noise and car noise at various SNR levels, namely, *10* dB, *5* dB and *0* dB. The performance evaluation for the various features in signal degradation conditions is shown in Table VI. The performance is shown in terms of % EER for vocoder-based (*S1-S9*) spoofs and overall % EER (*S1-S10*). For white noise, it is observed that the energy features gave almost equal average % EER in clean and in the presence of *10* dB and 5 dB noise. A similar case was observed with that of the shape and energy features fused at feature-level. The % EER of the shape and energy features, when used jointly, was more than energy features till *5* dB. However, for *0* dB SNR, the use of shape and energy features gave the best performance of *29.6%* EER. For these features, the spoof-specific information was preserved at severe signal degradation conditions as well. The overall % EER of the *8-D* shape and energy features seems to improve slightly with the signal degradation. However, at *10* dB, the % EER increases for *S1-S9* and decreases for *S10* spoof and hence, the overall % EER shown in Table VI decreases slightly. On the other hand, the frequency-domain features were severely affected by noise. Especially for MFCC feature set, at *0* dB SNR, the performance degrades to around *44%* with *D3* feature vector. The *FrFR1* and *FrFR5* features were found to perform better than MFCC features till *5* dB SNR noise.

For babble noise, the % EER for all features increases even for *10* dB SNR. The % EER is least for *3-D* energy features and is maximum for MFCC feature set. For *0* dB SNR, the least average EER is *28.6%* obtained using *FrFR5* feature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2682788, IEEE Journal of Selected Topics in Signal Processing

TABLE VI
% EER ON TESTING WITH THE EVALUATION SET USING THE SOURCE-FILTER INTERACTION BASED FEATURES AND MFCC AT VARIOUS FEATURE DIMENSIONS IN THE PRESENCE OF ADDITIVE WHITE NOISE, BABBLE NOISE AND CAR NOISE AT VARIOUS SNR LEVELS

| **Feature Sets** | | **Energy** | | **Shape+Energy** | | **FrFR1** | | | | | | **FrFR5** | | | | | | **MFCC** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dimensions** | | **3-D** | | **8-D** | | **D1** | | **D2** | | **D3** | | **D1** | | **D2** | | **D3** | | **D1** | | **D2** | | **D3** | |
| **% EER →**<br>**SNR (dB) ↓** | | **S1-S9** | **Avg.** | **S1-S9** | **Avg.** | **S1-S9** | **Avg.** | **S1-S9** | **Avg.** | **S1-S9** | **Avg.** | **S1-S9** | **Avg.** | **S1-S9** | **Avg.** | **S1-S9** | **Avg.** | **S1-S9** | **Avg.** | **S1-S9** | **Avg.** | **S1-S9** | **Avg.** |
| **Clean** | | 4.93 | 13.1 | 12.9 | 19.8 | 6.37 | 14.2 | 5.14 | 13.1 | 5.13 | 12.9 | 11.7 | 18.3 | 9.19 | 15.9 | 7.35 | 14.3 | 1.48 | 5.49 | 0.56 | 5.49 | 0.32 | 4.25 |
| **White** | **10** | 3.6 | 11.5 | 13.1 | 19.7 | 23.4 | 28.1 | 12.9 | 19.6 | 12.0 | 18.8 | 24.0 | 28.0 | 25.5 | 28.8 | 26.9 | 30.2 | 39.8 | 42.5 | 40.3 | 43.6 | 38.4 | 41.5 |
| | **5** | 8.3 | 13.6 | 14.5 | 19.2 | 29.4 | 33.8 | 21.1 | 27.4 | 18.6 | 24.9 | 34.7 | 36.9 | 30.9 | 33.1 | 32.9 | 35.2 | 41.6 | 44.1 | 48.1 | 51.5 | 40.9 | 43.2 |
| | **0** | 31.3 | 31.3 | 28.4 | 29.6 | 39.2 | 41.3 | 44.3 | 46.8 | 42.5 | 45.2 | 53.6 | 53.5 | 49.0 | 49.5 | 46.4 | 47.0 | 42.0 | 44.1 | 50.5 | 53.8 | 43.1 | 44.1 |
| **Average** | | **14.4** | **18.8** | 18.7 | 22.8 | 30.7 | 34.4 | 26.1 | 31.3 | 24.4 | 29.6 | 37.4 | 39.5 | 35.1 | 37.1 | 35.4 | 37.5 | 41.1 | 43.6 | 46.3 | 49.6 | 40.8 | 42.9 |
| **Babble** | **10** | 8.7 | 16.8 | 14.3 | 20.6 | 16.9 | 23.4 | 14.2 | 21.3 | 16.0 | 22.8 | 22.0 | 27.1 | 22.0 | 27.5 | 20.2 | 25.7 | 45.8 | 50.2 | 35.7 | 40.9 | 30.8 | 36.3 |
| | **5** | 12.9 | 20.1 | 19.5 | 24.5 | 25.9 | 30.8 | 22.6 | 28.2 | 25.3 | 30.3 | 29.2 | 33.2 | 29.9 | 34.3 | 27.8 | 32.0 | 45.6 | 49.6 | 45.0 | 49.0 | 40.2 | 44.0 |
| | **0** | 27.5 | 32.1 | 27.8 | 31.7 | 33.4 | 36.5 | 33.8 | 36.8 | 36.5 | 39.1 | 29.8 | 32.8 | 25.6 | 29.6 | 24.9 | 28.6 | 42.9 | 46.0 | 48.7 | 51.7 | 44.5 | 46.8 |
| **Average** | | **16.4** | **23.0** | 20.5 | 25.6 | 25.4 | 30.2 | 23.5 | 28.8 | 25.9 | 30.7 | 27.0 | 31.0 | 25.8 | 30.5 | 24.3 | 28.8 | 44.8 | 48.6 | 43.1 | 47.2 | 38.5 | 42.4 |
| **Car** | **10** | 2.7 | 11.6 | 15.0 | 22.0 | 7.8 | 15.6 | 7.3 | 15.4 | 7.4 | 15.3 | 17.2 | 22.8 | 14.1 | 20.3 | 12.6 | 19.0 | 20.0 | 26.7 | 9.6 | 16.1 | 15.1 | 22.8 |
| | **5** | 4.4 | 13.1 | 19.6 | 26.0 | 11.1 | 18.7 | 13.2 | 21.2 | 13.6 | 21.5 | 19.3 | 24.5 | 17.7 | 23.5 | 16.8 | 22.7 | 27.1 | 33.4 | 12.6 | 18.9 | 21.6 | 29.0 |
| | **0** | 13.6 | 20.7 | 22.3 | 27.9 | 17.1 | 23.6 | 19.3 | 26.0 | 20.7 | 27.5 | 29.8 | 32.8 | 25.6 | 29.6 | 24.9 | 28.6 | 34.6 | 40.1 | 15.9 | 21.8 | 28.4 | 35.2 |
| **Average** | | **6.9** | **15.1** | 19.0 | 25.3 | 12.0 | 19.3 | 13.3 | 20.9 | 13.9 | 21.4 | 22.1 | 26.7 | 19.1 | 24.5 | 18.1 | 23.4 | 27.2 | 33.4 | 12.7 | 18.9 | 21.7 | 29.0 |

set. The *FrFR1* and *FrFR5* representations, i.e., the S-F interaction cues in the frequency domain were able to better classify natural *vs.* spoofed speech in the presence of babble noise as compared to white noise. However, on an average of all SNRs, the *3-D* energy features perform the best amongst all the features.

Next, for car noise, the % EER for all the features (except *3-D* energy features) increased at *10* dB SNR. However, the performance degradation was less as compared to white and babble noise. Interestingly, at *10* dB SNR using *3-D* energy features, the % EER improved both for vocoder-based cases (*S1-S9*) as well as for the average % EER. This is because the % EER of vocoder-based *S7* spoof decreased about *10* times in the presence of *10* dB car noise. This was also observed for *10* dB white noise as well where the detection for *S7* and *S10* was improved. Similar observations were found in [51], where the performance on *S10* improved in the presence of reverberation noise. It was observed that with the temporal filtering of reverberation, the discontinuity in *S10* spoof could become more obvious. However, much needs to be explored about the improved performance even in the presence of noise.

In recent works, where white noise is considered [62]- [63], it was observed that the % EER at *0* dB was as high as *40%* obtained by fusing several features. In such multiple fusions of features, it is difficult to conclude why a particular feature performs well for a particular noise. In the present case, the energy features gave better performance even at *0* dB SNR. This is because, in the present approach, the excitation source features in the time domain are obtained by inverse filtering from the speech signal the high frequency resonances corresponding to the vocal tract system. Thus, the high frequency noise is also filtered out. Therefore, the shape parameters along with the energy-based features help in maintaining the performance of the SSD systems much better as compared to the MFCC features that contain much broader spectra.

Figure 12 shows the % EER for known attacks (*S1-S5*), unknown attacks (*S6-S10*), vocoder-based spoofs (*S1-S9*) and average EER (*S1-S10*) (averaged across the various SNR levels) for *3-D* energy features, *8-D* shape and energy features, *FrFR1*, *FrFR5* and MFCC feature sets for the *D3* dimension. It can be observed that MFCC is highly sensitive to any type of signal degradation conditions as compared to the frequency-domain features that use S-F cues. The simple approach of residual using energy-based features proves to be effective in the presence of noise as well without significant performance degradation. It is also observed that, on average, white and babble noise were more severe as compared to car noise and at low SNR values, the white noise affects the performance more than the babble noise.

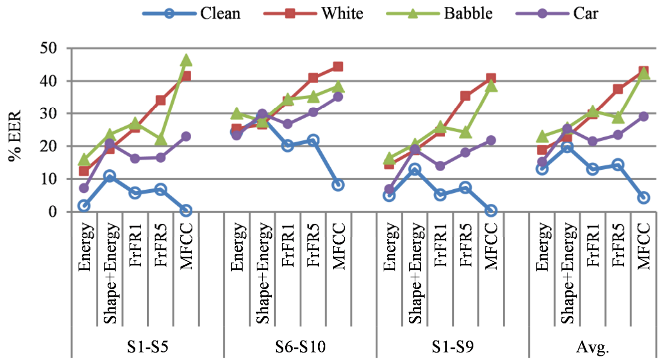

Fig. 12. The % EER for known attacks (*S1-S5*), unknown attacks (*S6-S10*), vocoder-based spoofs (*S1-S9*) and average EER (*S1-S10*) averaged across the various SNR levels for *3-D* energy features, *8-D* shape and energy features, FrFR1, FrFR5 and MFCC feature sets for *D3* dimension.

## *G. Results on the Blizzard Challenge 2012 Dataset*

To evaluate the channel mismatch case, the GMMs are trained on the ASV spoof 2015 data and tested on the Blizzard Challenge 2012 dataset. A complete representation of the time-domain shape and energy features and frequency-domain residual features at all feature vectors is shown in Fig. 13. As observed from the results, the problem of channel mismatch cannot be generalized based on the performance of countermeasures on the ASV spoof 2015 challenge database. Generally, the performance of MFCC decreases with the addition of Δ and ΔΔ features to the static features. However, in this case, the % EER for MFCC does not always decrease with the addition of dynamic features as compared to the gradual decrease in % EER by the *FrFR1* and *FrFR5* features. Thus, MFCC with the *D3* feature vector cannot be considered optimum in all the cases.

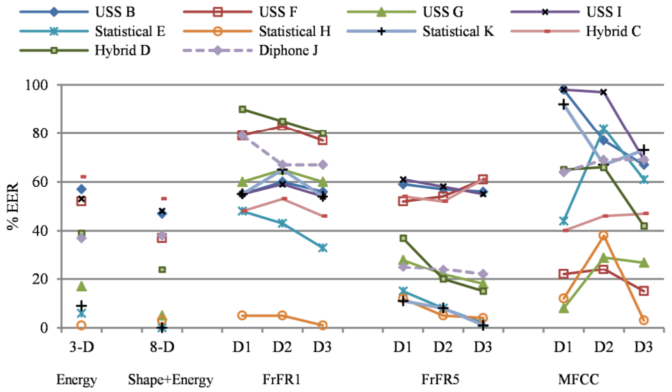

Fig. 13. The % EER on the Blizzard Challenge 2012 dataset for *3-D* energy features, *8-D* shape and energy features, *FrFR1*, *FrFR5* and MFCC feature sets at all feature vectors.

For the shape and energy features, the % EER is much less for statistical-based synthetic speech as compared to USS-based, hybrid or diphone-based synthetic speech. An average representation across USS, statistical, hybrid and diphone-based systems is shown in Table VII. For the frequency-domain features, the *D3* feature vector is considered. Among all the systems, the USS-based systems were found to be difficult to detect in the SSD task. In the case of SPSS-based systems, the shape and energy-based features gave much less % EER. For hybrid and diphone systems, the performance was similar to that of the USS-based speech. The *FrFR5* features detected hybrid and diphone systems with least 38% and 22% EER, respectively. On the whole, for the channel mismatch case, the performance of shape and energy features was found to be better than rest of the other features. The performance of all the feature sets, in this case, is highly dependent on the type of attacks. Therefore, there is a need to examine the type of training strategy used such that the channel variability can be handled.

TABLE VII

AVERAGE % EER ON TESTING WITH BLIZZARD CHALLENGE 2012 DATABASE FOR DIFFERENT FEATURES (TRAINING = ASV SPOOF 2015 DATABASE)

| Feature Sets→ | Energy | Shape + Energy | FrFR1 | FrFR5 | MFCC |
|---|---|---|---|---|---|
| Systems ↓ | 3-D | 8-D | D3 | D3 | D3 |
| USS | 44.8 | **34.25** | 61.8 | 47.5 | 44.5 |
| Statistical | 5.3 | **1.00** | 29.3 | 2.30 | 45.7 |
| Hybrid | 50.5 | 38.5 | 63.0 | **38.0** | 44.5 |
| Diphone | 37.0 | 38.0 | 67.0 | **22.0** | 69.0 |

## VI. SUMMARY AND CONCLUSIONS

This study presented the use of features motivated from the natural human speech production mechanism. Each time the vocal folds open and close, there exists a nonlinear source and system interaction. This interaction is also known to be present in the residual component $g_r(t)$. The feature extraction studied assumes that a genuine trial (i.e., natural utterance) includes cues caused by the interaction between the glottal source and vocal tract while these cues are absent from an impostor trial (i.e., an utterance produced by a text-to-speech system or by a voice conversion system). Thus, to summarize, the S-F interaction cues consist of estimating the glottal excitation source, $\dot{g}(t)$, followed by obtaining a parametric representation using the LF-model. Thereafter, the residual is computed by subtracting the LF-modeled coarse structure from the estimated flow. Finally, both time- and frequency-domain processing is used to express the extracted residual component as a feature vector to classify natural *vs.* spoofed speech. The nonlinear interaction is an attribute specific to the natural speech and is difficult to incorporate it in the machine-generated speech. We exploit this property and, hence, observe that for the vocoder-based spoofed speech, the results are very encouraging for the SSD task.

In this study, we presented the significance of the $R_d$ shape parameter of the LF-model in interpreting the characteristics or quality of the speech. For example, if HMM-based SS sounds breathy, then this is reflected in the shape parameter. However, not much can always be inferred about the naturalness of speech due to the fact that the vocoded speech is a subset of the natural speech. The shape parameters capture the entire shape of the glottal flow and hence, may embed information about the glottal skewness (which is mainly due to S-F interaction). However, this is not explicitly carried out in this work. In addition to the shape parameters, the residual energy in the closed, open and return phase is considered. It is known that the ripple component of the residue embeds information about the nonlinear interaction. This is because nonlinear interaction and modulations are possible only during the production of the natural speech and not the vocoded speech. For the vocoder-based speech (*S1-S9*), the *3-D* energy features gave *4.930%* and the *36-D* MFCC features gave *0.336%* EER. However, on the score-level fusion of the energy features with MFCC, the EER of MFCC features drops down by *10%*. Thus, the extensive analysis carried out in this study indicates that the use of residual energy-based features aids in the spoof detection task.

It has been observed that the higher frequency regions of speech are essential for the SSD task [23]. However, in the present case, the residual signal has a ripple or fine structure component with frequency around the first formant ($F_1$). The HFR features that are speaker-specific are suppressed due to inverse filtering of vocal tract information. In addition, to enhance the LFR, the Mel representation of the residual is taken. Furthermore, being in the frequency domain, this representation captures the residual information at a higher dimension than just using the average values in the closed,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2682788, IEEE Journal of Selected Topics in Signal Processing

open and return phase. Next, the residues in the frequency domain are indicative of the formant modulation in the frequency domain. These modulations are present in natural speech and not in vocoded speech. Therefore, considering S-F interaction features, the vocoded speech is detected better. That is, to the score-level fusion of energy and MFCC features, when the *FrFR1* features are again fused, the EER for vocoder-based *S1-S9* spoof drops by ~ *29%*.

The S-F interaction features represent information of the voice excitation source at a lower frequency region than the actual speech signal. Thus, this is indeed a promising approach as these features are likely to be robust under signal degradation conditions. As observed, these features perform better (even at lower SNRs) than MFCC or any such features that process the entire spectrum of the speech signal (i.e., available bandwidth for a particular sampling frequency) with average EER around *15-30%* on the evaluation set. Furthermore, the results of testing on a completely unrelated database such as the Blizzard Challenge 2012 showed that the time-domain excitation source features perform very well especially for the vocoder-based speech (*~1%* EER) and the performance is also better for the hybrid, diphone and USS-based systems as compared to the MFCC feature set. Spoof detection in signal degradation and channel mismatch conditions is an important research issue and needs further investigation as the features should generalize in terms of better performance for the clean speech as well.

Considering the literature of SSD, the use of S-F interaction features is not much yet explored. However, in this study, an underlying assumption is that the ripple component exists due to the S-F interaction only. There may be other reasons due to which the ripple may occur such as poor antiresonances due to biasing of the lowest formants by the harmonics and improper fitting of the coarse structure. It is assumed here that while using the inverse filtering technique, the formants are cancelled well in order to get appropriate inverse filtered speech to estimate the glottal flow waveform. Thus, the ripple estimation may depend on the choice of inverse filter (e.g., if the F1 is not properly cancelled either in the closed or the open phase). In addition, the proper fitting, i.e., of the coarse structure $g_c(t)$, may also affect the feature extraction process, and the system performance. Thus, the S-F interaction features could be further explored with respect to their issues and used for the SSD task. Our future research efforts will be directed to explore the proposed features for other signal degradation conditions such that they generalize for the clean speech as well. In addition, the use of LSFs of the excitation source, such as the glottal flow waveform or the residual signal, to capture distinct characteristics as compared to the LSFs of the vocal tract spectrum can also be explored. Furthermore, to detect concatenative speech synthesis, we plan to analyze the formant modulation transitions at the joints to develop SSD systems even for vocoder-independent synthesis systems.

## References

[1] Z. Wu, et al., "Spoofing and countermeasures for speaker verification: A survey," *Speech Comm.*, vol. 66, pp. 130-153, Feb. 2015.

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

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

[4] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Workshop on Speech Synthesis*, Santa Monica, CA, USA, 2002, pp. 227-230.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Comm.*, vol. 51, no. 11, pp. 1039-1064, Nov. 2009.

[6] J. Yamagishi, et al., "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 66-83, Jan. 2009.

[7] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131-142, Mar. 1998.

[8] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Transfer function-based voice transformation for speaker recognition," in *Proc. IEEE Speaker Lang. Reco. Workshop (Odyssey)*, San Juan, Puerto Rico, 2006, pp. 1-6.

[9] Z. Wu, et al., "SAS: A speaker verification spoofing database containing diverse attacks," in *Proc. IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Brisbane, Australia, 2015, pp. 4440-4444.

[10] Z. Wu, et al., "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2037-2041.

[11] 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. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2062-2066.

[12] J. Sanchez, et al., "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2042-2046.

[13] L. Wang, Y. Yoshida, Y. Kawakami and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2092-2096.

[14] J. Sanchez, et al., "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Trans. Info. Forensics and Security*, vol. 10, no. 4, pp. 810-820, April 2015.

[15] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. Int. Conf. on Spoken Lang. Process., (ICSLP)*, Beijing, China, 2000, pp. 302-305.

[16] A. Ogihara, H. Unno, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Trans. on Fund. of Elect., Comm. and Comp. Sciences*, vol. E88-A, no. 1, pp. 280-286, 2005.

[17] P. L. De Leon, B. Steward, and J. Yamagishi, "Synthetic speech discrimination using pitch patten statistics derived from image analysis," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Oregon, USA, 2012, pp. 370-373.

[18] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency ($F_0$) and strength of excitation (SoE) for spoofed speech detection," in *Proc. IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP)*, Shanghai, China, 2016, pp. 5105-5109.

[19] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2077-2081.

[20] H. Bhavsar, T. Patel, and H. Patil, "Novel nonlinear prediction based features for spoofed speech detection," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, San Francisco, USA, 2016, pp. 155-159.

[21] 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. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2082-2086.

[22] X. Xiao, et al., "Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2052-2056.

[23] M. Sahidullah, T. Kinnunen, and C. Hanilçi , "A comparison of features for synthetic speech detection," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2087-2091.

[24] H. Yu, et al., "Effect of multi-condition training and speech enhancement methods on spoofing detection," in *Proc. Int. Workshop on Sensing, Process. and Learning for Intelligent Machines (SPLINE)*, Aalborg, Denmark, 2016, pp. 1-5.

[25] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients," in *Proc. Odyssey 2016: The Speaker and Language Recognition Workshop,* Bilbao, Spain, 2016, pp. 283-290.

[26] 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. Int. Speech Comm. Assoc. (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2097-2101.

[27] M. H. Soni, T. B. Patel, and H. A. Patil, "Novel subband autoencoder features for detection of spoofed speech," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, San Francisco, USA, 2016.

[28] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoof detection," *Speech Comm.*, vol. 85, pp. 43-52, 2016.

[29] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson India, Eight Impression, 2012.

[30] P. M. Morse and K. U. Ingard, *Theoretical Acoustics*. Princeton University Press, 1987.

[31] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Labs Technical Journal (BLTJ)*, vol. 51, no. 6, pp. 1233-1268, July 1972.

[32] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *KTH Quaterly Progress and Status Report (STL-QPSR)*, vol. 26, no. 4, pp. 001-013, 1985.

[33] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *KTH Quaterly Progress and Status Report (STL-QPSR)*, vol. 36, no. 2-3, pp. 119-156, 1995.

[34] J. P. S. R. Cabral, "HMM-based speech synthesis using an acoustic glottal source model," Ph.D. Thesis, Centre for Speech Technology Research, University of Edinburgh, 2010.

[35] T. Raitio, et al., "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 1, pp. 153-165, 2011.

[36] T. Yoshimura, et al., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eur. Conf. Speech Process. Techno. (EUROSPEECH)*, Budapest, Hungary, 1999, pp. 2347-2350.

[37] T. Yoshimura, et al., "Mixed excitation for HMM-based speech synthesis," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Aalborg, Denmark, 2001, pp. 2263-2266.

[38] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Proc. of Workshops on Models and Analysis of Vocal Emissions for Bio. Appl. (MAVEBA)*, Firenze, Italy, 2001, pp. 1-6.

[39] Y. Stylianou, "Concatenative speech synthesis using a harmonic plus noise model," in *Proc. ESCA Speech Synthesis Workshop (SSW)*, Jenolan Caves House, NSW, Australia, 1998, pp. 261-266.

[40] 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 Lang. Process. (ICSLP)*, Yokohama, Japan, 1994, pp. 1043-1046.

[41] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. Int. Conf. on Acou. Speech and Sig. Process. (ICASSP)*, Pennsylvania, USA, 2005, pp. 9-12.

[42] L. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Comm.*, vol. 28, no. 3, pp. 211-226, 1999.

[43] M. D. Plumpe, "Modeling of the glottal flow derivative waveform with application to speaker identification," Masters Thesis, Dept. of Electrical Engg. and Comp. Science, MIT, Feb. 1997.

[44] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. on Speech and Audio Process.*, vol. 7, no. 5, pp. 569-586, Sept. 1999.

[45] I. Titze and A. Palaparthi, "Sensitivity of source-filter interaction to specific vocal tract shapes," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 24, no. 12, pp. 2507-2515, Dec. 2016.

[46] I. Titze, "Nonlinear source-filter coupling in phonation: Theory," *Jour. Acoust. Soc. Amer. (JASA),* vol. 123, no. 5, pp. 2733-2749, May 2008.

[47] I. Titze, T. Riede, and P. Popolo, "Nonlinear source-filter coupling in phonation: Vocal excercises," *Jour. Acoust. Soc. Amer. (JASA)*, vol. 123, no. 4, pp. 1902-1915, Apr. 2008.

[48] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Comm.*, vol. 1, no. 3-4, pp. 167-184, 1982.

[49] C. R. Jankowski, "Fine structure features for speaker identification," Ph.D. Thesis, Dept. of Electrical Engg. and Comp. Science, MIT, 1996.

[50] J. Kane, et al., "Exploiting time and frequency domain measures for precise voice source parameterisation," in *Proc. Speech Prosody*, Shanghai, China, 2012, pp. 143-146.

[51] X. Tian, et al, "An investigation of spoofing speech detection under additive noise and reverberant conditions," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, San Francisco, USA, 2016, pp. 1715-1719.

[52] I. Saratxaga, et al., "Synthetic speech detection using phase information," *Speech Comm.*, vol. 81, pp. 30-41, April 2016.

[53] Speech Synthesis Special Interest Group (SynSIG): Blizzard Challenge. [AvailableOnline].http://www.synsig.org/index.php/Blizzard_Challenge {Last accessed 09th August 2016}.

[54] S. King and V. Karaiskos, "The Blizzard Challenge 2012," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Portland, Oregon, USA, 2012, pp. 1-11.

[55] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse," *Speech Comm.*, vol. 11, no. 2-3, pp. 109-118, 1992.

[56] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. on Speech and Audio Process.*, vol. 16, no. 8, pp. 1602-1613, Nov. 2008.

[57] J. Kane. Voice_Analysis_Toolkit: A set of MATLAB codes for carrying out glottal source and voice quality analysis [Available Online]. https://github.com/jckane/Voice_Analysis_Toolkit. {Last accessed: 6th Dec. 2016}.

[58] D. Pati and S. R. M. Prasanna, "Processing of linear prediction residual in spectral and cepstral domains for speaker information," *Int J Speech Technol (IJST)*, vol. 18, pp. 333-350, 2015.

[59] T. B. Patel and H. A. Patil, "Cochlear filter and instantaneous frequency based features for spoofed speech detection," *accepted* in *IEEE Jour. of Selected Topics in Sig. Process., (JSTSP),* Special Issue on Spoofing and Countermeasures for Automatic Speaker Verification, Dec. 2016.

[60] A. Martin, G. Doddington, T. Kamm, and M. Ordowski, "The DET curve in assessment of detection task performance," in *Proc. Eur. Conf. Speech Process. Techno. (EUROSPEECH)*, Greece, 1997, pp. 1895-1898.

[61] ASVspoof 2017: Automatic Speaker Verification Spoofing and Countermeasures Challenge, Theme in the 2017 edition: Audio replay attack detection. [Available Online]. http://www.spoofingchallenge.org/

[62] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection under noisy conditions: A preliminary investigation and an initial database," http://arxiv.org/pdf/1602.02950v1.pdf, 2016.

[63] C. Hanilci, T. Kinnunen, M. Sahidullah, and A. Sizov, "Spoofing detection goes noisy: An analysis of synthetic speech detection in presence of additive noise," *Speech Comm.*, vol. 85, pp. 83-97, Dec. 2016.

Note: Photograph and brief bio-data of the authors is available in our companion paper in the same special issue.