Machine Learning Mitigants for Speech Based Cyber Risk

MARTA CAMPI\textsuperscript{1}, GARETH W. PETERS\textsuperscript{2}, NOURDDINE AZZAOUI\textsuperscript{3}, AND TOMOKO MATSUI\textsuperscript{4}, (Senior Member, IEEE)

\textsuperscript{1}Department of Statistical Science, University College London (UCL), Bloomsbury, London W1T 7PJ, U.K.
\textsuperscript{2}Department of Statistics and Applied Probability, University of California at Santa Barbara, Santa Barbara, CA 93106, USA
\textsuperscript{3}Laboratoire de Mathématiques Blaise Pascal, Université Clermont Auvergne, 63178 Aubière, France
\textsuperscript{4}Department of Statistical Modeling, The Institute of Statistical Mathematics, Tachikawa-shi, Tokyo 190-8562, Japan

Corresponding author: Marta Campi (marta.campi11@gmail.com)

This work was supported in part by The Institute of Statistical Mathematics, Tachikawa-shi, Tokyo, Japan.

ABSTRACT Statistical analysis of speech is an emerging area of machine learning. In this paper, we tackle the biometric challenge of Automatic Speaker Verification (ASV) of differentiating between samples generated by two distinct populations of utterances, those of an authentic human voice and those generated by a synthetic one. Solving such an issue through a statistical perspective foresees the definition of a decision rule function and a learning procedure to identify the optimal classifier. Classical state-of-the-art countermeasures rely on strong assumptions such as stationarity or local-stationarity of speech that may be atypical to encounter in practice. We explore in this regard a robust non-linear and non-stationary signal decomposition method known as the Empirical Mode Decomposition combined with the Mel-Frequency Cepstral Coefficients in a novel fashion with a refined classifier technique known as multi-kernel Support Vector machine. We undertake significant real data case studies covering multiple ASV systems using different datasets, including the ASVSpoof 2019 challenge database. The obtained results overwhelmingly demonstrate the significance of our feature extraction and classifier approach versus existing conventional methods in reducing the threat of cyber-attack perpetrated by synthetic voice replication seeking unauthorised access.

INDEX TERMS Cyberattack, empirical mode decomposition, support vector machine, speech analysis.

I. INTRODUCTION

The prevalence of biometric authentication systems is increasing in many data access points in smart devices and remote data access settings. This has led to a new machine learning based approach to address the resulting biometric challenge of Automatic Speaker Verification (ASV). Modern machine learning approaches are recently tackling the study of ASV, see [1] and [2]. In this paper, we also look at a novel machine learning solution for ASV that is designed around feature extraction for speech signals, and we address the challenge of biometric cyber-attack mitigation by seeking to detect when data access is attempted through a deep fake artificial speech generation rather than a human speaker. In the same vein as the biometric verification work for fingerprints of [3], we will be performing identification and verification speech biometrics.

The key statistical component of our proposed speech signature representation is based upon a non-stationary functional basis characterisation for speech signal via the Empirical Mode Decomposition (EMD). The EMD [4] is a basis decomposition method dealing with non-stationary and non-linear signals. It dynamically decomposes a signal into oscillatory, locally adapting AM-FM (amplitude and frequency modulated) components [5], called Intrinsic Mode Functions (IMFs). We employ the EMD to identify which voice signal components provide discriminatory power in mitigating the risk associated with biometric cyber attacks in ASV technology frameworks, where the extracted IMF basis functions act as an individual’s vocal signature allowing for discrimination of the human voice from synthetic attacks using replicated artificial voice. The EMD has been employed within speech analysis in [5]; while [6] made use of the EMD for the noise-robustness of automatic speech recognition systems. Reference [7] focuses on speech-based emotion classification utilising acoustic data and successfully employed the EMD basis functions and the instantaneous frequencies derived...
through the Hilbert transform. Furthermore, [8] used the EMD algorithm to extract the fundamental frequency $F_0$.

ASV technologies are gaining widespread utilization in contexts of call centers, human-computer interfaces, secure access control for commercial and retail banking, see [9], and [10]. An ASV system typically extracts speaker characteristics from utterances and compares them to a given speaker synthetic voice model, estimated from its identity. In this context, one may distinguish between text-dependent and text-independent frameworks. The former uses a fixed collection of reference sentences, while the latter exploits purely arbitrarily selected speech utterances. These are usually referred to as Text-Dependent Speaker Verification systems, or TD-SV, versus Text-Independent systems, or TI-SV (see discussions in [9]). A further differentiation might be given by speaker-dependent verification systems (SD-SV) or speaker-independent verification systems (SI-SV), where the former are trained by the individual who uses the system, while the latter trained as a system-agnostic to who is then using it. As with any biometric system, ASV is subject to spoofing or presentation attacks, which mimic a target speaker’s voice in person or remotely via artificial tools such as voice conversion (VC) or speech synthesis (SS) algorithms. The study of such attacks is of growing significance in the services industry, particularly the financial services sector, where clients’ data access is increasingly reliant on biometric identification. Spoofing attacks on banking records may be classed as a form of cyber attack. We seek to provide a machine learning classifier solution to detect and mitigate losses to data integrity and sensitive information by detecting and preventing synthetic voice access attacks.

Consequently, a range of approaches is emerging to produce specific countermeasures to mitigate against different types of cyber spoofing attacks (see [10], [11] for ASV and [12] for a survey focusing on speaker recognition presenting several countermeasures). The standard approach in many of these countermeasures is to identify speech parameterisations carrying discriminative power to differentiate between spoofed and real voices. The designed techniques make use of a classifier that attempts to distinguish between samples from two distinct populations of utterance, those from authentic voice and those from a synthetic generation of voice, derived from the two classes of speech signals [13]. The raw speech time-domain signals are often transformed into lower-dimensional sets of summary statistics or engineered feature representations for such classifiers, see [14]. Furthermore, such countermeasures often rely on standard time-frequency techniques constrained by assumptions such as stationarity or linearity of the underlying speech signal. The speech community has proposed multiple variations of these classical methods to overcome the aforementioned issues (see for example [15]–[18], [19], [20]) and so dealing with different aspects faced by ASV systems in discriminating spoofed and real voices. The traditional practice foresees the extraction or engineering of the raw speech data features and then conducts the classification task by stacking them within a vector. In this way, the classifier is often polluted by multi-frequency content information all contained in the proposed unique vector. The approach proposed in this work aims to tackle such a problem by constructing a parsimonious model that separates this frequency information content instead and selects the most discriminant areas of the time-frequency plane regarding the speech scenario analysed.

A recently developed approach dealing with ASV challenges is given by Deep Learning (DL). The reader should refer to [21] for an overview of deep learning based speaker recognition approaches that could also be extended to speaker verification tasks. These include multi-stage networks, end-to-end networks, generative networks or meta-learning. As highlighted in [21], these techniques are at a stage of minimal investigation with no asserted guidance on how to perform them efficiently and compare them to existing methodologies. Furthermore, DL requires, in general, high computational costs associated with big data training sets, often making them difficult to use in practice, and, therefore, further research is required to establish this direction. In the specific setting of ASV challenges, the idea behind DL methods, particularly the one of Deep Neural Network (DNN) quickly becoming the “new-state-of-the-art” methods, is to identify the formants structure with the complex function using many layers of perceptrons. This procedure is a high-cost learning procedure that will be replaced in this work by the EMD technique, able to capture formant structure with the requirement of much fewer parameters and can be applied to small and large datasets providing a uniform method in this regard. Hence, a sparse architecture in the placement of DNNs is promoted by this work. Afterwards, a much simpler classifier relying on the recent method known as multi-kernel Learning [22] combined with the Support Vector Machine is proposed.

Another important aspect is that not only speech is highly non-stationary and non-linear per se, but when ASV challenges are solved, adverse environments might be the one of interest, making the task even more difficult. For example, the presence of noise affecting human speech during the recording or the need for a very long speech signal to be recorded by the user to train the system or reverberation affecting the system. These challenges are discussed in [23], where the authors propose a method for short fragments of speech signals tackling the issues above described.

Given the great variety of approaches introduced in the literature and the several databases built and considered by researchers, it is hard to identify a uniform, standard technique unifying the presented framework and tackling the explained issues. The desired and sought technique should carry three main properties: first, non-stationarity and non-linearity should be heavily considered since speech is highly affected by these two characteristics. Often, real-world settings can be further corrupted by adverse environments such as noise, which can cause classical Fourier methods to fail to provide reliable and consistent results across different experiments and noise environments. Secondly, the
discrimination power of the classifier should be the centre of attention, and new classification methodologies should be proposed and studied. Third, several benchmark ASV features have been proved to be successful in multiple cases. The focus should be on the statistical interpretability of the ones able to identify discriminating insights in solving the ASV challenge.

Our approach follows along the familiar line of attack mitigation adopting a classifier framework, and the novelty lies in three components: the ability to treat the feature extraction in a non-stationary formulation; secondly, an ensemble learning multi-kernel classification framework is developed, see [22]; thirdly, the interpretability of the given feature extraction framework in terms of formant structures differentiating real and spoofed speech is provided. We demonstrate that it can improve the ability to detect attacks when compared to current state-of-the-art methods.

Furthermore, three datasets will be considered for the conducted experiments setting up multiple speech scenarios as text-dependent or text-independent and speaker-dependent and speaker-independent. Among these datasets, two of them are constructed explicitly by the authors without a recording laboratory or particular microphones providing the setting encountered in ASV challenges of adverse environments. Therefore, the obtained results will provide robustness in these settings.

A. CONTRIBUTIONS AND NOVELTY

The contributions of this work involve several core elements: firstly, enhanced non-stationary time-frequency methods applied to perform novel feature extraction techniques for the capture of speech signatures or vocal fingerprints. Secondly, using these new feature extraction methods to formulate a multi-kernel classifier based on Support Vector Machine techniques. This is highly beneficial in the classification tasks depending on the speech system considered: the extracted features are often combined in a unique vector, and the SVM is then performed. Such a practice should be avoided since it will add noise to the classification problem mixing the formant structure depending on both the individual and the gender. Therefore, if the analysed scenario is text-dependent or text-independent or, for example, speaker-dependent or speaker-independent, the standard operation of considering a unique feature vector characterising the entire time-frequency plane would pollute the classification learning procedure. The third contribution foresees the performance comparisons between benchmark ASV features extracted on the raw data and on the EMD basis functions to highlight that speech is highly non-stationary and that multiple situations generate adverse environments that require the use of an adaptive method relying on the given data system. Afterwards, the proposed methodology is tested through the use of various TTS algorithms within different speech scenarios. To achieve such a goal, we developed the following components:

1) We extend existing speech engineering techniques to non-stationary basis extraction methods and re-express them within a statistical framework. This is achieved via Empirical Mode Decomposition methods, which we use to extract time-domain intrinsic mode basis functions, which we represent via semi-parametric spline model characterizations.

2) The instantaneous frequency of each Intrinsic Mode basis function is derived in closed form via the Hilbert transform analytic extension. We are then able to apply the combination of time-domain non-stationary basis characterization of the speech signals and the instantaneous frequency characterizations to form a complete time-frequency signature of a person’s vocal and speech characteristics. We demonstrate that such basis functions are more amenable to classical speech feature extraction methods in the transformed cepstral domain. This allows us to develop new approaches to EMD-Mel Cepstral speech signature characterization that we demonstrate is highly effective in capturing individual speakers vocal tract specificities that arise given a speakers glottal airflow shaped by the vocal tract filter as it passes through it to produce speech. We can then use these features to distinguish between real speech and artificial computer-generated spoofed synthetic speech by capturing these signature features.

3) Our resulting speech signature feature characterizations allow us to solve important new biometric tasks related to detecting cyber intrusion attempts to access biometrically multi-factor secured data or systems where speech is one of the security factors. We have developed a class of multi-kernel support vector machine classifier solutions to detect such cyber attacks attempted through synthetically generated speech. These contributions then form a complete system, summarised in Figure 1, for a cyber threat detection framework capable of accurately detecting synthetic spoofed voice attacks on a speech based biometric system secure access.

B. BACKGROUND ON STATISTICAL CHARACTERIZATION OF SPEECH SIGNALS

According to the source-filter model [24], a speech signal is the result of the glottal airflow shaped by the vocal tract filter as it passes through it [25]. Under such a representation,
it is common to consider two main feature classes for an ASV system: voice source features or vocal tract features. The former are indeed related to the source of voiced sounds deriving from the glottal flow; however, numerous studies provide evidence showing that vocal folds features are not as discriminatory as vocal tract features [26]. We, therefore, focus our attention on the vocal tract features and in particular representations that contain information about the resonance properties of the vocal tract, also known as formants.

An individual’s speech formant structures are analogous to that individual’s speech fingerprint, thereby characterizing unique traits of the filter model specific to a human. Such features are, therefore, highly discriminatory, as it is challenging for a synthetic voice model to capture these individual-specific characteristics, see [27]. Considering features that can capture information on formant structures is crucial to mitigate biometric speech attacks on ASV-based security systems successfully.

Several methods can be employed to try to extract aspects of formant feature information, and they are often based on basis decomposition techniques, see [5], [28]. Such methods aim to separate the signal into components whose frequency spectras could be preferably dominated by a single non-overlapping formant frequency. All such methods currently work under assumptions of stationarity; for instance, see the Linear Prediction (LP) analysis framework [29]. However, as demonstrated in [29], such analysis provides an inaccurate estimation of the vocal tract resonances and the excitation source of the speech signal. A widely used alternative is to adopt warped filter basis extraction methods applied to windowed raw speech signal segments. A popular choice in practice is the Mel Frequency Cepstral Coefficients (MFCCs), see [30]. In this work, we intend to demonstrate that utilising non-stationary basis representations for speech enhances the ability to identify formant structures. This will be achieved by developing the Empirical Mode Decomposition (EMD) basis representations, see [4]. We develop a novel framework that utilises the EMD to define adaptive non-stationary features which efficiently detect highly frequent temporal variations characterising the original speech time-series. The EMD features are further combined with MFCCs so that summaries of speech capturing intrinsic non-stationarity and formant structures are contemporaneously detected. Related approaches have been considered mixing these concepts; we cite amongst other [31] and [32]. In the former, the authors propose the EMD as a dyadic filter in substitution to the mel-filter banks commonly used for the MFCCs. The extracted coefficients were therefore filtered according to the EMD basis. In [32], authors compute the MFCCs for the speech signal, and, after, the EMD is calculated for each coefficient. Our approach differs from both since the Mel Frequency Cepstral Coefficients are performed to represent the extracted non-stationary EMD basis themselves. We argue this will outperform alternative methods since it removes the requirement of local stationary assumptions that the methods mentioned above required for the first stage of the MFCC transforms. The traditional assumption made in speech is that speech signals should be approximately stationary at 25 milliseconds sampling rate under ideal background noise conditions. However, ASV systems would often operate within non-ideal environments affected by background noise or interference, which will be captured along with voices (see [33] and [34]). Instead, the EMD basis functions will accommodate non-stationarity of any level and so produce more robust features.

C. NOTATION

The following notation is used throughout: \( \gamma_k(t) \) represents the \( k \)-th Intrinsic Mode Function of the EMD basis functions; \( K \) is the total number of convexity changes of the original signal; \( f_s(t) \) is the instantaneous frequency of the \( k \)-th IMF; \( \gamma_k(t) \) analytic extension of \( \gamma_k(t) \); \( j \) is the complex unit; \( \mathcal{H}\{\cdot\} \) is the Hilbert transform; \( \mathcal{H}\{\cdot\} \) represents the Shannon entropy; \( \mathcal{H} \) represents the Hilbert space of transformed features; \( \omega \) is the raw data sampling frequency (Hz); \( \omega \) is the sampling frequency of the constructed IMFs (Hz) (according to Nyquist rule); \( \phi \) is the Mel-scale frequency; \( \mathcal{m}(l) \) is a Mel Frequency Cepstral Coefficient; \( k(x, y) \) represents a kernel function; \( \mathcal{K} \) represents the Gram Matrix associated to a kernel function; \( \langle \cdot, \cdot \rangle \) is the inner product; \( |\cdot| \) represents the cardinality set; \( || \cdot || \) is the indicator function.

II. EMPIRICAL MODE DECOMPOSITION FOR NON-STATIONARY FEATURES

The Empirical Mode Decomposition (EMD) non-stationary basis extraction approach is not widely known to the statistical audience; consequently, we will review a few core concepts of the EMD. Assume we have observed a continuous non-stationary speech signal \( s(t) \) through a sample recording at times \( 0 = t_1 < \cdots < t_K = T \). When applying the EMD basis decomposition framework, we first convert the partially observed discrete signal \( s(t) \) into a continuous analog signal, which we denote by \( \tilde{s}(t) \). To achieve this we use natural cubic polynomial splines, we will also then express the EMD bases \( \{\gamma_k(t)\}_{k=1}^{K} \) as natural cubic splines, derived from representation \( \tilde{s}(t) \).

Definition 1: Given a set of \( l \) knots \( a = t_1 < t_2 < \cdots < t_l = b \), a function \( \tilde{s} : [a, b] \rightarrow \mathbb{R} \) is called a cubic polynomial spline if:

- \( \tilde{s}(\cdot) \) is a polynomial of degree 3 on each interval \( (t_j, t_{j+1}) \) \( (j = 1, \ldots, l - 1) \)
- \( \tilde{s}(\cdot) \) is twice continuously differentiable

It is then a natural cubic spline when \( \tilde{s}'(a) = \tilde{s}'(b) = 0 \).

Hence, the speech signal representation \( \tilde{s}(t) \) is expressed in the class of truncated power basis, where the knot points are placed at the sampling times \( (t_i = t_i) \)

\[
\tilde{s}(t) = a_0 + a_1t + a_2t^2 + a_3(t - t_1)^3_+ + \cdots + a_{l-2}(t - t_{l-1})^3_+.
\]
The coefficients are estimated by standard penalised least squares

\[
\sum_{i=1}^{N-1} (s(t_i) - \tilde{r}(t_i))^2 + \lambda \int_{t_1}^{t_{N-1}} \tilde{s}''(t)^2 \, dt
\]

(2)

with natural cubic spline constraints \( \tilde{s}''(0) = \tilde{s}''(t_N) = 0 \) and where \( \lambda > 0 \) controls smoothness of the representation (see [35]). In this case, the number of total convexity changes (oscillations) of the analog signal \( \tilde{s}(t) \) within the time domain \([0, t_N]\) is denoted by \( K \in \mathbb{N} \). One may now define the EMD decomposition of a speech signal \( \tilde{s}(t) \) as follows.

**Definition 2:** The Empirical Mode Decomposition of signal \( \tilde{s}(t) \) is represented by the finite number of non-stationary basis functions known as Intrinsic Mode Functions (IMFs), denoted by \( \{\gamma_k(t)\} \), such that

\[
\tilde{s}(t) = \sum_{k=1}^{K+1} \gamma_k(t) + r(t)
\]

(3)

where \( r(t) \) represents the final residual (or final tendency) extracted, which has only a single convexity. Remark that the above decomposition can be written as:

\[
\tilde{s}(t) = \sum_{k=1}^{K} \gamma_k(t)
\]

(4)

where, for simplicity, the residual \( r(t) \) is denoted as \( \gamma_{K+1}(t) \). We employ such a notation within the section VII. In general the \( \gamma_k \) basis will have \( k \)-convexity changes throughout the domain \((t_1, t_N)\) and each IMF satisfies:

- **Oscillation** The number of extrema and zero-crossing must either equal or differ by most at one by:

\[
\text{abs}\left(\left\{\left\{\frac{d\gamma_k(t)}{dt} \right\} = 0 : t \in (t_1, t_N)\right\}\right) \] - \( \{\gamma_k(t) = 0 : t \in (t_1, t_N)\}\] \in \{0, 1\}

(5)

- **Local Symmetry** The local mean value of the envelope defined by a spline through the local maxima denoted \( \tilde{z}^{U_k}(t) \) and the envelope defined by a spline through the local minima denoted by \( \tilde{z}^{L_k}(t) \) is equal to zero pointwise i.e.

\[
m_k(t) = \frac{(\tilde{z}^{U_k}(t) + \tilde{z}^{L_k}(t))}{2} I(t \in [t_1, t_N]) = 0
\]

(6)

The minimum requirements of the upper and lower envelopes are:

\[
\tilde{z}^{U_k}(t) = \gamma_k(t), \quad \text{if} \quad \frac{d\gamma_k(t)}{dt} = 0 \quad \text{and} \quad \frac{d^2\gamma_k(t)}{dt^2} < 0,
\]

\[
\tilde{z}^{L_k}(t) \geq \gamma_k(t) \quad \forall t \in (t_1, t_N)
\]

\[
\tilde{z}^{L_k}(t) = \gamma_k(t), \quad \text{if} \quad \frac{d\gamma_k(t)}{dt} = 0 \quad \text{and} \quad \frac{d^2\gamma_k(t)}{dt^2} > 0,
\]

\[
\tilde{z}^{L_k}(t) \leq \gamma_k(t) \quad \forall t \in (t_1, t_N).
\]

(7)

Further, note that, in the above representation, \( \gamma_k(t) \) is not explicitly expressed in a functional form, as opposed to classical stationary methods where a parametric family of basis functions are stated, such as a cosine basis or a wavelet basis function. Here, the basis can take any functional form so long as it satisfies the decomposition relationship and the properties stated for an IMF. We utilise throughout the same flexible natural cubic spline representation as used to represent the speech signal \( \tilde{s}(t) \).

Note that each IMF carries a unique number of convexity changes that can occur at any time spacings. Typically, the times of convexity change are irregularly spaced and reflect non-stationarity in a local bandwidth of the frequencies that characterise the signal at that time instant. As a result of this property, one can still order the basis IMF’s naturally according to the unique number of total convexity changes they produce in \((t_1, t_N)\).

As outlined in [4], the construction of an IMF basis is directly linked to the concept of local symmetry required to handle non-stationary data. This notion is enclosed by the mean envelope that captures a local time scale, and the definition of a local averaging time scale is hence bypassed. Such a requirement is fundamental to avoid asymmetric waves affecting the concept of instantaneous frequency, formalised below.

**A. EXTRACTION OF EMD BASIS FUNCTIONS (IMFs)**

We next briefly outline the process applied to extract the IMF basis representations recursively. This procedure is known as **sifting**. The first step consists of computing extrema of \( \tilde{s}(t) \); this can be done based on observations or on the interpolated signal, \( \tilde{s}(t) \). Using \( \tilde{s}(t) \), the roots of the first derivative \( \tilde{s}'(t) \) produce the sequence of time points for successive maxima and minima:

\[
\left\{\tau_j^*\right\}_{k=1}^{K} = \left\{t \in [t_1, t_N]| a_1 + 2a_2 t + 3 \sum_{i=3}^{3+2j-2} a_i (t - \tau_i)^2 = 0\right\}.
\]

(8)

Without loss of generality, we assume the maxima occur at odd intervals, i.e. \( \tau_{2j+1}^* \), and minima occur at even intervals, i.e. \( \tau_{2j}^* \).

The second step of sifting builds an upper \( (\tilde{z}^{U_k}(t)) \) and lower \( (\tilde{z}^{L_k}(t)) \) envelope of \( \tilde{s}(t) \) using two natural cubic splines through the sequence of maxima and the sequence of minima respectively:

\[
\tilde{z}^{U_k}(t) = a_{0}^{U_k} + a_{1}^{U_k} t + a_{2}^{U_k} t^2 + \sum_{i=0}^{[K/2]} a_{i+3}^{U_k} (t - \tau_{2i+1}^*)^3_+,
\]

\[
\tilde{z}^{L_k}(t) = a_{0}^{L_k} + a_{1}^{L_k} t + a_{2}^{L_k} t^2 + \sum_{i=0}^{[K/2]} a_{i+3}^{L_k} (t - \tau_{2i}^*)^3_+,
\]

(9)

such that \( \tilde{z}^{U_k}(\tau_{2j+1}^*) = \tilde{s}(\tau_{2j+1}^*) \) for all odd \( \tau_{2j}^* \) and \( \tilde{z}^{L_k}(\tau_{2j+1}^*) \geq \tilde{s}(t) \) and equivalently \( \tilde{z}^{L_k}(\tau_{2j}^*) = \tilde{s}(\tau_{2j}^*) \) for all even \( \tau_{2j}^* \) and \( \tilde{z}^{L_k} \leq \tilde{s}(t) \). One then utilises these envelopes to construct the mean
signal denoted by $m_k(t)$ given in equation (6), which will then be used to compensate the original speech signal $\tilde{s}(t)$ in a recursive fashion, until an IMF is obtained. These bases are recursively extracted, this means that, once the $k$-th IMF is computed, it is subtracted from the main signal and the sifting procedure is applied to the residual signal to obtain the next IMF which will have one less convexity change than the previously extracted IMF on $(t_1, t_N)$. The procedure is detailed in Algorithm 1 in the Supplement Materials and stopping criteria are discussed in [36]. We illustrate the sifting process for IMF basis extraction in Figure 2.

![Figure 2](image.png)

**FIGURE 2.** First two steps of sifting for an IMF $\gamma(t)$.

### B. Obtaining Instantaneous Frequency From IMF Basis Functions

Classical Fourier methods require stationarity, where the frequencies of base components are pure harmonics that are static over time [4]. Real world signals such as in speech analysis are often non-stationary and non-linear and, therefore, carry time-varying frequency components. The EMD basis functions, IMFs, will admit a time-varying frequency structure that can be characterized by instantaneous frequencies (IFs). The IF of a given IMF basis is extracted in the following steps. First one takes the Hilbert Transform of each $\gamma_k(t)$, so that we can construct an analytic extension. The Hilbert Transform can be computed in closed form readily if $\gamma_k(t)$ respects the restrictions defined in (7). Define the analytic signal $z_k(t) = \gamma_k(t) + j\tilde{\gamma}_k(t) = a_k(t)\exp(j\theta_k(t))$. The analytic extension of $\gamma_k(t)$ with time varying amplitude $a_k(t) = \sqrt{\tilde{\gamma}_k^2(t) + \gamma_k^2(t)}$ and time varying phase $\theta_k(t) = \text{arctan}\frac{\tilde{\gamma}_k(t)}{\gamma_k(t)}$. Then $\tilde{\gamma}_k(t)$ is obtained via Hilbert Transform as follows:

$$\tilde{\gamma}_k(t) = \frac{1}{\pi} \lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} \frac{\gamma_k(t)}{t - \tau} d\tau$$  \hspace{1cm} (10)

Once the analytic extension is defined and, therefore, $z_k(t)$ is obtained, then the instantaneous frequency $f_k(t)$ for IMF $k$ is defined as:

$$f_k(t) = \frac{1}{2\pi} \frac{d\theta_k(t)}{dt} = \frac{1}{2\pi} \frac{\tilde{\gamma}_k(t)\gamma_k(t) - \tilde{\gamma}_k(t)\gamma_k(t)'}{\gamma_k^2(t) + \tilde{\gamma}_k^2(t)}$$  \hspace{1cm} (11)

We see that [4] imposed the conditions (7) characterizing the IMFs properties to then ensure that the instantaneous frequency remains positive and therefore admits a meaningful physical interpretation.

It will be advantageous to obtain the Hilbert transform of the $k$-th IMF by considering the natural cubic spline representation per knot segmentation as a local cubic polynomial for $t \in [t_{i-1}, t_i]$. Then the Hilbert transform is constructed as the following sum of local cubic polynomial transforms, see for details [37]:

$$\hat{\gamma}_k(t) = \mathcal{HT}\{\gamma_k(t)\} = \frac{1}{\pi} \sum_{i=1}^{N-1} \hat{\gamma}_k(t) - \gamma_k(t_{i-1} < t \leq t_i)$$  \hspace{1cm} (12)

where $\Delta_i = t_i - t_{i-1}$ and $\hat{\gamma}_k(t)$ is the Hilbert transform of the $i$-th polynomial:

$$\hat{\gamma}_k(t) = \left( a_{k1}t^3 + b_{k2}t^2 + c_{k3}t + d_{k4} \right) \log\left( \frac{t}{t - \Delta_i} \right)$$

$$+ a_{k5} \left( \frac{\Delta_i^2}{2} - \Delta_i^2 / 3 \right)$$

$$+ b_{k6} \left( -\Delta_i^2 / 2 - c_{k7} \right) - c_{k8}\Delta_i.$$  \hspace{1cm} (13)

### C. Interpreting EMD Basis Decomposition

Having extracted the IMFs via the sifting process, then for each IMF, one can evaluate the analytic extension using the Hilbert transform. This allows one to obtain a signal representation for $\tilde{s}(t)$ expressed in a “Fourier-like” expansion as:

$$\tilde{s}(t) = \text{Re} \left\{ \sum_{k=1}^{K+1} a_k(t)\exp(j\theta_k(t)) \right\}$$

$$= \text{Re} \left\{ \sum_{k=1}^{K+1} a_k(t)\exp(j\int_{t_1}^{t} 2\pi f_k(t)dt) \right\}$$  \hspace{1cm} (14)

in which the residual $r(t)$ is set as the $K + 1$-th term. In the case that the target signal is made of a finite number of pure stationary harmonics as shown in Figure 3 denoted $s_1(t)$ and $s_2(t)$ respectively, the IMF decomposition will match the finite collection of Fourier bases as shown. When the signal is not comprised of a finite number of pure harmonics or is non-stationary, then the instantaneous frequencies for the IMF bases are not pure harmonics. However, the IMF bases extracted from EMD sifting decomposition can still be naturally ordered, but in a different manner to classical notions of frequency orders in Fourier analysis. They are ordered by oscillation count (total convexity changes) rather than frequency. This is not equivalent as the IMF bases are not, in general, strictly periodic. Due to this interesting difference, the IMFs may have some time intervals in $(t_1, t_N)$ where a high order IMF may have lower instantaneous frequency than...
FIGURE 3. Top panels represent $\tilde{s}(t) = \sin(4\pi t) + \sin(10\pi t)$. Bottom panels provide the two IMFs basis functions.

III. EMD-MFCC SPEECH SIGNATURES VIA PITCH AND VOCAL RESONANCE

In speech analysis, the formant frequencies act like a characteristic signature of a given speakers vocal tract, like a speech fingerprint that is characteristic of given speakers vocal tract physiology, see [38], [39]. Formants are a concentration of speech acoustic energy, usually occurring at approximately each 1,000Hz frequency band, directly related to the oscillatory modes of resonance of an individual vocal tract structure. They are often indexed by $F_1$, $F_2$, $F_3$, etc., where $F_0$ is termed the fundamental frequency and represents the rate at which the vocal folds vibrate. This quantity corresponds to the pitch and coincides with the first harmonic, $H_1$; harmonics are multiple of the fundamental frequency $F_0$ characterizing the glottal source. Suppose one can extract these features from non-stationary voiced speech created by a human vocal, physical, physiological system. In that case, they may have the potential to be highly discriminatory factors to distinguish a human versus a synthetic voice as they represent how vocal tracts shape sound sources which therefore have representations unique to an individual.

These formant features are often approximated by a Mel Cepstral basis projection, where the functional coefficients form the MFCC representation of the speech signal that approximate the formants. Such a characterization is working well in capturing formant structure in ideal speech recording environments with sufficient sampling rates to capture local stationary approximations of the non-stationary speech signal. However, in real-world ASV systems that we consider, speech is recorded in noisy real-world environments with more compressive sampling rates and background non-stationary noise and distortions. The presence of background noise and distortion have been shown in [33] to render the MFCC estimated coefficients as highly sensitive and not statistically robust to a variety of potential types of background noise and distortions. Furthermore, the compression of the signal prior to transmission to the ASV for comparison in the biometric signal analysis can further create aliasing distortions.

We will overcome these challenges by merging EMD with MFCC, where rather than passing the raw speech signal into the MFCC representation, we will first decompose the speech signal into IMF basis representations, then we will perform MFCC representations of each IMF basis as illustrated in Figure 5. This can be shown to robustly estimate the formant structures even in the presence of different speech signal recording distortions and background noise environments.

FIGURE 4. Top panel: signal $\tilde{s}(t) = \sin(4\pi t) \cdot [t \leq t_1] + \sin(15\pi t) \cdot [t > t_1]$. Middle panel: IMF extracted to represent $s(t)$ and Bottom panel: instantaneous frequency for IMF.

FIGURE 5. Diagram of our proposed methodology characterising EMD-MFCC features for formant detection.

There are existing works that have explored the development of EMD methods to characterize formant structures, see [40]. However, as explained in [5] they suffer from an identification complication known as mode-mixing, which is the inability to align formant structures and IMFs. This occurs since these previous works apply the EMD method to signals already based on stationary Fourier transforms of the non-stationary speech signal. In our work, we avoid the problem of mode-mixing by first performing the EMD basis decomposition of the speech signal, then we study the Mel Frequency Cepstral Coefficients (MFCCs) of each IMF basis. In this way, we can align exactly the formants with the ordering of the IMF bases represented through a second stage MFCC family of coefficient functions. The MFCC acts as a warped linear filter for each IMF expressed through a functional coefficient in time and fixed local frequency
selective basis. The resulting coefficients of the filter will be non-linearly spaced in their spectral energy so that they can be estimated to align with standing wave patterns of pitch and harmonics of human speech formants.

We define the MFCC representation as follows, starting from the base Mel-scale:

\[ \phi = 2595 \log_{10} \left( 1 + \frac{\tilde{\omega}}{700} \right), \]

where \( \phi \) is the subjective pitch in Mels corresponding to the original frequency \( \tilde{\omega} \) in Hz [41]. Let us consider the \( k \)-th IMF \( \gamma_k(t) \) extracted from speech signal representation \( \tilde{s}(t) \). Next, we provide a representation of the EMD-MFCC characterisation we propose, followed by a brief numerically stable approximation that also works well in practice. We first pre-emphasise and Hamming-windowed, \( \gamma_k(t) \) to get \( \gamma_k^*(t) \) to guard against issues of aliasing in discrete sample MFCC representations of each IMF basis.

We then decimate the continuous signal \( \gamma_k^*(t) \) to a set of \( T_s \) evaluated “sample” values in the local window frame. We then obtain a discrete vector representation \( \gamma_k^* = \{ \gamma_k^*(\frac{\tilde{\omega}_h}{T_s}) \}_{h=1}^{T_s-1} \) for \( \tilde{\omega}_h \) sampling frequency in Hz. Then perform the spectral transform of the \( k \)-th IMF representation \( \gamma_k^* \) to obtain local Fourier representation \( \Gamma_k^* \) given by DFT as:

\[ \Gamma_k^*(h) = \sum_{n=0}^{T_s-1} \gamma_k^*(t) \exp(-j2\pi \rho_h \tilde{\omega}_n / T_s) \]

The magnitude of spectrum \( |\Gamma_k^*(h)| \) is then scaled in both frequency and magnitude. The frequency is scaled through convolution with a linear Mel filter bank \( H(h, m) \), a multiplicative transfer function in the frequency domain, and then the logarithm of the result is taken to stretch or time-dilate the resulting signal. The output of this process is a collection of functional Mel Cepstral Coefficients for the \( k \)-th IMF given in the frequency domain by,

\[ \mathcal{M}_k(m) = \log_{10} \left( \sum_{h=0}^{T_s-1} PM_k^*(h) \right) \]

\[ = \log_{10} \left( \sum_{h=0}^{T_s-1} |\Gamma_k^*(h)| \cdot H(h, m) \right) \]

for \( m = 1, 2, \ldots, M \), where \( M \) is the number of Mel bases used (or order of the filter bank). The Mel filter bank is a sequence of triangular basis defined by the center frequencies \( \tilde{\omega}_m \) as follows:

\[ H(h, m) = \begin{cases} 0 & \tilde{\omega}(h) < \tilde{\omega}_m(m-1), \\ \tilde{\omega}_m(m-1) - \tilde{\omega}_m(m-1) & \tilde{\omega}_m(m-1) \leq \tilde{\omega}(h) < \tilde{\omega}_m(m), \\ \tilde{\omega}_m(m) - \tilde{\omega}_m(m+1) & \tilde{\omega}_m(m) \leq \tilde{\omega}(h) < \tilde{\omega}_m(m+1), \\ \tilde{\omega}_m(m) - \tilde{\omega}_m(m+1) & \tilde{\omega}(h) \geq \tilde{\omega}_m(m+1), \end{cases} \]

which satisfies \( \sum_{m=1}^{M} H(h, m) = 1 \). The center frequencies of the basis are computed through equation (15) to approximate the Mel scale. Afterwards, a fixed frequency resolution of the Mel scale is computed, which is a logarithmic scaling of the repetition frequency, obtained by \( \Delta \phi = (\phi_{\text{max}} - \phi_{\text{min}})/(M + 1) \), where \( \phi_{\text{max}} \) and \( \phi_{\text{min}} \) are computed with equation (15) by using \( \tilde{\omega}_{\text{max}} \) and \( \tilde{\omega}_{\text{min}} \) respectively and \( M \) is the number of basis (filter banks). The center frequencies on the Mel scale are given by \( \phi_{\text{c}}(m) = m \cdot \Delta \phi \) for \( m = 1, 2, \ldots, M \). In order to obtain such center frequencies, the inverse of equation (15) is used and then they are substituted in 18 to obtain the Mel filter banks. The Mel Basis is illustrated in Figure 6 for 40 filter banks with a sampling frequency of 44.1 kHz giving 1102 samples, which is the one used in our real speech data case study. Note that the higher is the frequency, the wider the filter banks become.

\[ \text{FIGURE 6. The Mel filter bank structure for 40 filters. Each peak represents the center frequency of the filters.} \]

We then convert the frequency domain Mel Cepstral functional coefficients back to the time domain via Discrete Fourier transform, which simplifies to a simple Discrete Cosine Transform (DCT) of \( \mathcal{M}_k(m) \) to obtain:

\[ r_k(l) = \sum_{s=1}^{M} \mathcal{M}_k(s) \cos \left[ \frac{l \pi}{M} \left( s - \frac{1}{2} \right) \right] \]

for \( l = 1, 2, \ldots, M \), where \( r_k(l) \) is the \( l \)-th MFCC of the \( k \)-th IMF.

Typical values for \( M \) in speech applications involve selecting the first 10-30 lowest center frequency cepstral coefficients; we, therefore, used 12 coefficients (the lowest) to model the individual speakers and the synthetic voice.

IV. CLASSES OF FEATURES TO REPRESENT SPEECH SIGNALS

The first step in a classification task is to summarise the underlying signals through feature extraction. The EMD captures several non-stationary attributes of the considered speech time-series through different spaces such as parameter space, basis space, instantaneous frequencies. This section aims to explain the representations we considered in ASV speaker verification applications.

In the table 1 we provide a summary of the multiple feature representations considered. The time mesh defined to summarise the features is denoted by \( t_i \) such that \( t_i \in \{0 = t_i^0, \ldots, t_i^N = N\} \). In obtaining the IMF and instantaneous frequency features, the EMD sifting process is performed
over each interpolated voice sample, and then 5 IMFs are stored, which capture a range of high frequency and low frequency structure: the first three with the highest number of oscillations, the lowest oscillation count and the residual with just one convexity sign. We will refer to such IMF basis functions either as EMD-MFCC or IMF-MFCC.

The goal is determining a hyperplane of separation with the maximum distance to the closest points of the identified classes. These points are called Support Vectors. By considering a training set \( \{(x_i, y_i)\}_{i=1}^{N} \), a feature vector \( x_i \in \mathbb{R}^D \) and class labels \( y_i \in \{-1, +1\} \), the hyperplane of separation can be defined as \( d(x_i, w, b) = w \cdot x_i + b = 0 \), where \( w \in \mathbb{R}^D \) represents the weight vector and \( b \) a scalar and the operation is a dot product. The optimal hyperplane that separates data into two classes is the one that minimises the following objective function:

\[
\frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i(w \cdot x_i - b)) + \lambda \|w\|^2. \quad (20)
\]

This corresponds to a quadratic optimisation problem and can be solved in the parameter space with respect to \( w \) and \( b \). There are several solutions to solve this problem, such as sub-gradient descent and coordinate descent methods. Starting from this primal form of optimization we next introduce slack variables where for all \( i \in \{1, \ldots, n\} \), the slack variables \( \xi_i = \max(0, 1 - y_i(w \cdot x_i - b)) \), measure the distance \( \xi_i \) between a point and its crossing of the margin. This relaxation allows one to accommodate less than perfect linear separation of training set data. The optimisation problem is then given by:

\[
\text{minimize } \frac{1}{n} C \sum_{i=1}^{n} \xi_i + \lambda \|w\|^2 \quad (21)
\]

subject to \( y_i(w \cdot x_i - b) \geq 1 - \xi_i \), and \( \xi_i \geq 0 \), for all \( i \). Note that \( C \) is the trade-off factor, compromising between the maximization of the margin and the minimization of the misclassification error.

The primal problem is typically reformulated to as a dual problem through a Lagrangian, and the solution is guaranteed if the Karush-Kuhn-Tucker conditions [43] are verified. By solving for the Lagrangian dual, the problem then becomes:

\[
\text{maximize } f(\alpha_1 \ldots \alpha_n)
\]

\[
= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) \quad (22)
\]

subject to \( \sum_{i=1}^{n} \alpha_i y_i = 0 \), and \( 0 \leq \alpha_i \leq \frac{1}{2n\lambda} \) for all \( i \). Since the dual maximization problem is a quadratic function of the \( \alpha_i \) subject to linear constraints, it is efficiently solvable by quadratic programming algorithms. Here, the variables \( \alpha_i \) are defined such that \( w = \sum_{i=1}^{n} \alpha_i y_i x_i \). Moreover, \( \alpha_i = 0 \) exactly when \( x_i \) lies on the correct side of the margin, and \( 0 < \alpha_i < (2n\lambda)^{-1} \) when \( x_i \) lies on the margin’s boundary. It follows that \( w \) can be written as a linear combination of the support vectors. The offset, \( b \), can be recovered by finding an \( x_i \) on the margin’s boundary and solving \( y_i(w \cdot x_i - b) = 1 \iff b = w \cdot x_i - y_i \). (Note that \( \hat{y}_j = y_i \) since \( y_i = \pm 1 \).)

The presented framework provides a linear classifier assuming linear separability of the data which is, in practice,
rare to observe. The solution tackling this problem is known as kernel trick and extends such methods to non-linear set-

ings by projecting the feature data $x_i \in \mathcal{X}$ (in Table 1) into a transformed feature space through a non-linear map $\phi(x_i)$ which if selected adequately will provide close to perfect linear separability. This map $\phi : \mathcal{X} \to \mathcal{H}$ is called the feature map, $\mathcal{H}$ is the transformed feature space. In most cases, knowledge of this mapping is difficult to select explicitly to achieve this objective, so instead it is common to utilise an implicit solution to selecting this mapping by replacing it with a kernel representation. Consider a kernel function defined on the original feature space, $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_\mathcal{H}$. It is selected to satisfy that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ acts as an inner product representation on the implicit Separable Hilbert Space $\mathcal{H}$ and feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x_i, x_j \in \mathcal{X}$ we have $k(\cdot, \cdot) = \langle \phi_n(x_i), \phi_n(x_j) \rangle_\mathcal{H}$. The classification vector $w$ in the transformed space satisfies $w = \sum_i^m \alpha_i y_i \phi(x_i)$ where, the $\alpha_i$ are obtained by solving the maximization problem with cost function $f$ given as follows:

$$f(\alpha_1 \ldots \alpha_n) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j k(x_i, x_j) \alpha_i \alpha_j$$

subject to $\sum_{i=1}^n \alpha_i y_i = 0$, and $0 \leq \alpha_i \leq \frac{1}{\gamma} \forall i$. The coefficients $\alpha_i$ can be solved for using quadratic programming, as before. Again, we can find some index $i$ such that $0 < \alpha_i < (2n\gamma)^{-1}$, so that $\phi(x_i)$ lies on the boundary of the margin in the transformed space. Finally, the optimal decision function of a classifier is

$$z \mapsto \text{sgn}(w \cdot \phi(z) - b) = \text{sgn}\left(\sum_{i=1}^n \alpha_i y_i k(x_i, z) - b\right),$$

producing non-linear classification decision boundaries dependent on the kernel choice and kernel hyper-parameters.

A. FAMILIES OF KERNELS AND MULTI KERNEL COMBINATION

It is widely known in machine learning that kernel algebra allows one to construct kernel functions in various means. For instance through a sequence of composite maps, where $k(x_i, x_j) = k(\bar{k}(\phi_{\delta}(x_i), \phi_{\delta}(x_j)))_\mathcal{H}$ is still a kernel when $\phi_{\delta}(x_i) = (\phi_{\delta} \circ \phi_{\delta-1} \circ \ldots \circ \phi_1)(x_i)$. In this paper, this is analogous to transforming the speech signal through multiple representations such as IMF decomposition, followed by MFCC characterisation. Alternatively, weighted linear combinations of kernels will also produce a valid kernel. This would be equivalent to combining feature maps, one from each IMF, one from each instantaneous frequency signal, one from each MFCC coefficient function, etc. This concept approach to multiple feature combining has become known as multiple kernel learning (MKL). The motivation behind these methods is the additional flexibility within the learning process and the need for representing multiple, heterogeneous data properties, see [22]).

In constructing such MKL frameworks, two main approaches can be considered: a two-stage process which learns first each kernel hyper-parameters and then secondly the combining function/weights. The other method jointly learns both the kernel hyper-parameters and the combining function/weights. In practice, it is common to encounter the use of a convex weighted combining rule

$$k_{\eta}(x_i, x_j) = \sum_{m=1}^M \eta_m k_m(x_i^m, x_j^m)$$

with $\eta_m \in [0, m]$ and $\eta_1 + \ldots + \eta_M = 1$. Within such a construction, each $k_m(x_i^m, x_j^m)$ characterizes a distinct sub-set of features of the data. It is then possible to interpret the contribution of each individual component to the learning process. The $\eta$ coefficients can be interpreted to understand which features are more relevant for discrimination. In order to estimate such $\eta$ weights, we adopt the approach of [44]; using the performance obtained by each kernel separately, select

$$\eta_m = \frac{\pi_m - \delta}{\sum_{m=1}^M (\pi_m - \delta)}$$

where $\pi_m$ is the accuracy of $k_m$ used individually and $\delta$ is the threshold that should be less than or equal to the minimum of the accuracies obtained from single-kernel learners. This work explores six different kernels as outlined in Table 2.

| Kernel  | Formula                                                                 | Optimization Parameter |
|---------|-------------------------------------------------------------------------|------------------------|
| RBF     | $k(x_i, x_j) = \exp(-\alpha \|x_i - x_j\|^2)$                         | $\alpha$, $\sigma$, $\gamma$ |
| Laplace RBF | $k(x_i, x_j) = \exp(-\alpha \|x_i - x_j\|)$                          | $\alpha$, $\sigma$      |
| Polynomial | $k(x_i, x_j) = (\alpha (x_i \cdot x_j) + \gamma)^d$                | $\alpha$, $\gamma$, $d$ |
| Sigmoid  | $k(x_i, x_j) = \text{tanh}(\alpha (x_i \cdot x_j) + \gamma)$         | $\alpha$, $\gamma$      |
| Bessel   | $k(x_i, x_j) = \frac{\text{Bessel}_d^{\alpha}(\sqrt{\|x_i - x_j\|^2})}{\|x_i - x_j\|^d}$ | $\alpha$, $\gamma$, $d$ |
| Linear   | $k(x_i, x_j) = (x_i \cdot x_j)$                                        | $\alpha$, $\gamma$, $d$ |

Successful SVMs are strictly dependent on the selected hyperparameters of the kernel functions. Optimal selections can be made for performance measurements evaluated through a cross-validation score of the training set. Several methods are available to search for optimal hyperparameters; the grid-search results tend to be the most numerically stable and easy to implement. In this work we set hyperparameter regions as follows: $C \in \{2^{-2}, 2^{-1}, \ldots, 2^6\}$, $\gamma \in \{2^{-6}, 2^{-4}, \ldots, 2^{-2}\}$, $d \in \{1, 2, 3\}$, $\nu \in \{1, 2\}$. Regarding the grid for $\alpha$, we adopted the kernlab package approach for R, which uses the sigmoid function to obtain the grid range for this parameter. The selected values for $\alpha$ corresponds to a trimmed mean of its grid. The case study is implemented with 2-fold cross-validation of the training set to tune hyperparameters.
VI. STATISTICAL PERSPECTIVE OF EMD-SVM FROM A DECISION-THEORY VIEW

This section provides two different interpretations to statistically interpret the Support Vector Machine: a loss function regularized settings and a Bayesian binary decision framework. Whilst this is known, we believe the reader should benefit from such an interpretation and better understand the proposed approach we adopted as part of our solution.

A. INTERPRETATION THROUGH REGULARISED LOSS FUNCTION

The SVM has become an important method in classification problems, yet it is often presented from an optimisation perspective. In this section, we provide a statistical perspective based on [45] and [46]. Supervised learning algorithms, such as the SVM, are given a set of training samples \( x_1, \ldots, x_n \) with labels \( y_1, \ldots, y_n \) to predict \( Y_{n+1} \) given \( X_{n+1} \). These methods make use of a hypothesis \( f \) such that \( f(X_{n+1}) \) will be an approximation of \( Y_{n+1} \). To achieve a reliable approximation, a loss function \( \ell(y, z) \) is associated with the risk of \( f \) measuring how different \( z \) is as a prediction of the true \( y \). We would then like to choose a hypothesis that minimises the expected risk, defined as the expectation of the loss function:

\[
e( f ) = \mathbb{E}[\ell(Y_{n+1}, f(X_{n+1}))]
\]  

(27)

In most cases, the risk \( e(f) \) cannot be obtained since the joint distribution of \( X_{n+1}, Y_{n+1} \) is unknown. A common strategy is choosing the hypothesis that minimizes an estimation of \( e(f) \) via the empirical risk:

\[
\hat{e}(f) = \frac{1}{n} \sum_{k=1}^{n} \ell(y_k, f(x_k))
\]  

(28)

Under certain assumptions about the sequence of random variables \( X_k, y_k \) (for example, that a finite Markov process generates them), if the set of hypotheses is small enough, the minimiser of the empirical risk will closely approximate the minimiser of the expected risk as \( n \) grows large.

In order for the minimisation problem to have a well-defined solution, we have to place constraints on the set \( H \) of hypotheses. If \( H \) is a normed space (as is the case for SVM), a particularly effective technique is to consider only those hypotheses \( f \) for which \( \| f \|_H < k \). This is equivalent to imposing a regularisation penalty \( \mathcal{R}(f) = \lambda_k \| f \|_H \), and solving the new optimization problem given as:

\[
\min_{f \in \mathcal{F}} \hat{e}(f) + \mathcal{R}(f) = \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{k=1}^{n} \ell(y_k, f(x_k)) + \lambda_k \| f \|_H
\]  

(29)

where \( \lambda \) gives the degree of the penalization, \( \mathcal{F} \) a decision function class, \( \ell \) represents the considered loss function and \( \| f \|_H \) is usually referred to as regularisation functional. Amongst other classifiers, we consider the large margin classifier Support Vector Machine, which uses the hinge loss, or soft margin loss, defined as \( [1 - yf(x)]_+ \).

B. INTERPRETATION THROUGH BAYESIAN DECISION PERSPECTIVE

A second statistical perspective on the SVM in the binary classification context is to consider such classifier as the solution to the classical binary hypothesis testing problem, in which a realisation \( x \) of the random variable \( X \) is observed from an observation space \( \mathcal{X} \), such that \( X \in \mathcal{X} \). There are two hypothesis labelled \( H_0 \) and \( H_1 \) defining a decision region \( \mathcal{X}_1 \in \mathcal{X} \) which rejects the null \( H_0 \) in favour of \( H_1 \) if and only if \( X \in \mathcal{X}_1 \). The test could be alternatively formalized through the binary value of \( \phi(X) \), where \( \phi \) represents the indicator function on \( \mathcal{X}_1 \). To interpret an SVM classifier as an inference framework, we may consider a Bayesian setting in which \( X \) follows a distribution \( \pi_0 \) under \( H_0 \) and \( \pi_1 \) under \( H_1 \). The log-likelihood ratio is defined by the logarithm of the Radon–Nykodim derivative as \( L = \log(d\pi_1/d\pi_0) \). Given a threshold \( c \in \mathbb{R} \), the log-likelihood ratio test (LRT) states \( H_1 \) to be true if and only if \( L(X) \geq c \). That means \( \mathcal{X}_1 = \{ x \in \mathcal{X} : L(x) \geq c \} \), and therefore, \( \phi(x) = \mathbb{I}(L(x) \geq c) \). Such decision rule (LRT) is taken into account since it reaches the minimum probability of error, called Bayes error, and it represents the best error rate that any classifier could achieve. An alternative perspective, achieving perfect discrimination is via the SVM; suppose there exists a set \( \mathcal{X}_1 \) such that \( X \in \mathcal{X}_1 \) under \( H_1 \) and \( X \in \mathcal{X}_1^c \) otherwise. To obtain an effective test, assume a given family of functions \( \mathcal{F} \); a test is then sought among the class of indicators \( \phi(x) = \mathbb{I}(f(x) \geq c) \), with \( c \) scalar threshold. As in [46], the next optimization problem is considered to construct a test that is optimal over this class:

\[
\Delta^* = \max_{f \in \mathcal{F}} \inf_{x_1, x_0} (f(x_1) - f(x_0))
\]  

(30)

where the infimum is over all training data \{\( x_1 \) \} observed under \( H_1 \) and \{\( x_0 \) \} observed under \( H_0 \). If \( \Delta^* > 0 \), then a maximizer \( f^* \) will produce a test that perfectly discriminates. In such cases, it is possible to conclude that for some \( c \in \mathbb{R} \):

\[
f^*(X) \geq c \text{ under } H_1 \text{ and } f^*(X) \leq c \text{ under } H_0
\]  

(31)

Such a criterion then finds its explanation in the LRT. Suppose that perfect discrimination is possible: let us take any pair of probability distribution \( \pi_0 \) and \( \pi_1 \) mutually singular and model data assuming \( X \sim \pi_i \) under \( H_i \). Then (31) is satisfied by using the LRT:

\[
L(X) = +\infty \text{ under } H_1 \text{ and } L(X) = -\infty \text{ under } H_0
\]  

(32)

It is precisely in this way, that the SVM optimisation problem can be enclosed within a statistical inference context.

Looking at SVM from this perspective is particularly helpful for classifier evaluations; this paper tries to differentiate between synthetic and real voices by setting up a binary classification that exploits several features. In binary classification problems, such as ours, assessing the ability of the classifier (SVM) is commonly done through the contingency table or confusion matrix. Such a table is built by comparing the true values of the two classes (previously obtained by labelling data as positive and negative values) and the predicted one computed by the classifier.

VOLUME 9, 2021
136841
VII. EXPERIMENTAL SET-UP

There are three classes of experiments shown in table 4 that foresee the use of three datasets, which are described in table 3. We introduce the three datasets, and then the different types of experiments are presented.

We first focus on datasets one and two since they consider two specific classes of sentences, respectively. We highlight the use of these two sets to test our novel methodology within a text-dependent and a speaker-dependent verification system (TD-SD-SV) relevant to ASV challenges characterised by these conditions. The first dataset involves a set of sentences constructed to be challenging and reflect a real ASV setting in which sentences are not phonetically balanced. We obtained them from the first text (Inferno) that makes up Dante Alighieri “The Divine Comedy”. The second dataset is a reference set based on the IEEE Recommended Practices for Speech Quality Measurements, as described in [47], extensively used in speech analysis testing of speaker verification. It sets out seventy-two lists of ten phrases described as the 1965 Revised List of Phonetically Balanced Sentences, otherwise known as the ‘Harvard Sentences’. These are widely used in telecommunications, speech, and acoustics research, where standardised and repeatable speech sequences are needed.

| TABLE 3. Description of the datasets employed throughout the various sets of experiments. The number of utterances for each speaker is balanced across each dataset. For example, in dataset one, training set, there are 800 utterances; given that the number of speakers is 8, this means 100 utterances per speaker. This is valid for every other set. For the classification tasks, gender has been taken into account. Hence, the speakers have been divided between male and female voices. The considered methodology aims to detect the energy concentration of the formant structure, which heavily differs amongst these two categories. Each dataset is further described within the text. The procedure applied to extract a subset of the ASVspoof 2019 challenge dataset is presented in VII-C. |
|---|---|---|---|---|
| Dataset | # of Speakers | # of Utterances | System |
| | Natural | Spoofer | Training | Testing |
| 1 | Not Phonetically Balanced | 2 | 6 | 800 | 160 | TD-SD-SV |
| 2 | Phonetically Balanced | 2 | 6 | 576 | 576 | TD-SD-SV |
| 3 | Subset of ASVspoof 2019 | 28 | 84 | 10160 | 2464 | TI-SI-SV |

In both datasets, two real-language sources were used from a female (speaker 1) and a male (speaker 2); for the synthetic speech, five correspondent sources (T1, T2, T3, T4, T5 described in table 10) were employed for the female case and one source (T1) for the male one. The synthetic speech voices of all TTS algorithms were selected to have an English accent. The voice recordings were sampled at 44.1kHz without significant channel or background noise to develop a text-dependent scenario relevant for speaker verification tasks [48]. Recording environments of both training and testing voice samples were identical to avoid mismatched conditions (see [13], and [48]). Common sentences were used for each speaker and the synthetic voice.

Note that no recording laboratory or specialised microphone was used, and the utterances were recorded in noisy, reverberant environments. This is particularly relevant since it sets up the setting for adverse environments commonly encountered in ASV challenges. Therefore, the obtained results will carry the added feature of robustness to these kinds of speech settings.

The duration of each sentence speech recording was approximately 15sec to 1min maximum producing between 661k and 2,646k samples per spoken sentence. The start and end of each sample were trimmed to remove any non-speech segments and decimated to a set of 60k total samples. Regarding the IMFs extraction procedure, each set of 60k samples for one sentence was then windowed into non-overlapping collections of 5,000 samples and passed to the EMD sifting procedure. Afterwards, the features presented in Table 1 were extracted. We note that in some cases, we found that for high-frequency instantaneous frequency features, it would be advantageous also to apply a median filter (we used a window of 2ms).

In the first dataset, the total number of recorded sentences was 960, equally proportioned samples of the same sentences across all voice recordings, with 80% randomly selected for training and the rest for testing. In the second dataset, we use the first sentence from each of the seventy-two lists of the Harvard Sentences to construct the training dataset. The testing dataset was given by the second sentence from each of the seventy-two lists of the Harvard Sentences. This led to 1,152 utterances split equally between training and testing sets.

The third dataset corresponds to a subset of the ASVspoof 2019 challenge database described in [55]. Details on the extracted sets of sentences are given in subsection VII-C. We underline the importance of the settings provided by this dataset: they will tackle text-independent and speaker-independent verification systems (TI-SI-SV) to test our novel methodology in the most general environment encountered in ASV challenges.

Table 4 presents the set of experiments considered. With experiment one, we firstly present a discussion showing how the EMD-MFCC approach provides more powerful discrimination in detecting individual vocal tracts required in ASV systems compared to other sets of EMD features (IMFs, IFs, Spline Coefficients, etc.). We also provide the benchmark model comparison of the traditional MFCC extraction on the raw speech signals presented in table 6. Furthermore, given the wide variety of features often employed in Speaker Verification or Speaker Recognition tasks (see [56]), we also propose additional benchmark features applied both on the raw data and on the IMFs. Table 5 provides a detailed description of such features, with the used configuration and references required for further understanding. Results of these features run on the IMFs are provided in table 7. We discuss results for the individual speech features and then introduce the EMD-MFCC-MKL framework.
Experiment two focuses on a different aspect often faced by ASV systems: the different TTS algorithms. Several techniques produce a spoofing attack: impersonation, synthetic speech or TTS, voice conversion, and replay. In this work, we only consider TTS spoofing attacks. As highlighted in their work, [57] explains how TTS algorithms can nowadays produce high-quality voice through several kinds of methods as concatenative TTS unit selection [58], statistical parametric TTS [59], formant synthesis [60] and Deep Learning-based procedures (see [61]–[64]). Each of these procedures carries specific pros and cons, highlighted in Figure 12. Note that [57] also suggest the hybrid approach as presented in [65]. We select the best performing features for dataset one and dataset two for the female voice only obtained in Experiment one and repeat a similar exercise by considering the different TTS algorithms presented in Table 10. We show the best performance and provide the additional results within the Supplement Materials.

Experiment three runs the EMD-MFCC-MKL solution for ASV systems on a selected subset of the ASVspoof 2019 challenge dataset. In this way, a text-independent and speaker-independent environment is tested. Results will be presented for a range of different TTS algorithms and male and female voices. We focus on the best performing cases and present the additional results in the Supplement Materials.

In each experiment, we focus on presenting key aspects of the out-of-sample analysis that represent the most challenging cases for assessing our proposed EMD-MFCC methodology. All additional results are provided in the Supplement Materials, all code and data sets, including user guides, are provided at https://github.com/mcampi111/Speech-Experiment.

A. EXPERIMENT ONE: BIOMETRIC CYBER RISK MITIGATION VIA SYNTHETIC VS REAL VOICE DISCRIMINATION

Throughout these sections, we focus on the female voice examples to present the results. We found they generally presented the more challenging task in TD-SD-SV scenarios given the wider variation in spectral energy in the speech signals and higher non-stationary generally present in the
formant structures in the 5kHz to 20kHz range. Note that results for the individual SVMs of the male voice are presented in the Supplement Materials.

We start by assessing the ability of the benchmark ASV features in the classification of real and synthetic voices. We extract such features for the female voice versus the synthetic voice generated with TTS algorithm T1 for dataset one. This is done firstly on the raw speech data, and then the features are extracted not on the raw speech but rather on the IMF basis function representations of the speech. It is this combined EMD-feature representation that we advocate as a framework able to significantly enhance the discriminatory power of each of the familiar spectral, temporal speech features. We show that performances are improved universally by adopting our proposed approach of EMD-features compared to just features on raw speech. Indeed, applying gold standard (Short-Time Scale Discrete Fourier Transform) ST-DFT based feature on the EMD functions produces greater discriminatory power since the EMD non-stationary bases are better adapted to the speech recording environment and the non-stationary nature of the speech signal. Amongst the selected benchmark ASV features, we show that MFCCs are the best performing and select those to construct our new methodology combining EMD-MFCCs. We provide further evidence showing how this method better captures the formant structure of a given speaker through spectrograms and other plots below presented.

Hence, the baseline reference to our proposed methodology will be the benchmark ASV features and, in particular, the MFCCS constructed from the raw speech data. This contrasts with our proposed methodology of first extracting the IMF bases and applying MFCC to each IMF to produce more significant discrimination. We argue that the non-stationarity of higher frequency components in speech is more pronounced than lower-frequency components. Consequently, low-frequency bandwidths should be more comparable in terms of relative performances between MFCCs and EMD-MFCC features. At these frequencies, the fundamental frequency $F_0$ more closely reflects a stationary component, and therefore, MFCCs should be equally performing over either method. The majority of the difference is expected at higher frequencies, where it is more likely that non-stationarity will be non-uniformly distributed. We highlight that, in this first experiment, the setting foresees a text-dependent and a speaker-dependent scenario. Hence, only one speaker at a time is considered for the classification task, and speakers are making use of the same sentences. This is highly relevant since when multiple speakers and utterances are considered, the expected results will change. This will be presented in experiment three.

### SVM WITH SPEECH BENCHMARK ASV FEATURES

We first set up the results for the benchmark comparison, which is based on applying the benchmark ASV features to the raw speech signal. Table 6 presents the results. Note that, for this task, we focus on the female voice discrimination task with TTS algorithm T1 for dataset one. The configuration applied to obtain such coefficients are presented in table 5. We performed one SVM per individual coefficient. We selected $M = 12$ as is the standard recommendation when utilizing these features in speech analysis. We present results for the radial basis function kernel. Other kernels have been employed and produced similar results.

The features are divided into cepstral coefficients or linear prediction coefficients, and they use various filter banks when the former are considered, different transforms are applied once the linear prediction coefficients are obtained when it comes to the latter ones. These variants find their roots within different purposes. MFCCs, for example, try to capture formants by mimicking the human cochlear auditory capacity; the LFCCs are a similar feature, making use of a linear filter bank rather than the mel-filter bank to obtain a higher frequency resolution at high frequency [66]. BFCCs represent an alternative to MFCCs [67] whose filter banks should replicate the basilar membrane placed inside the cochlea that contains sensory receptors for hearing and performs spectral analysis for speech intelligibility perception. GFCCs [68] make use of the Gammatone filter bank for their cepstral analysis and model physiological changes in the inner ear and external, middle ear. IMFCCs consider the inverted-mel-filter bank and give a high-frequency resolution to low frequencies rather than high frequencies. We also consider MSRCCs and PSRCCs proposed in [19] whose final goal is to model the human auditory system by the functional relationship between the onset firing rate of auditory neurons and sound pressure level. The former ones capture information about the magnitude spectrum while the latter ones about the phase spectrum.

The NGCCs [20] use a Normalized Gammatone filter bank and incorporates the properties of the peripheral auditory system aimed to improve robustness in noisy speech settings. We then propose linear prediction coefficients and variations as the LPCCs, the PLPs and the RPLPs. These features rely on the stationarity of the underlying system, and even framing the speech signal into batches at which it turns stationary does not tackle the issue, especially when adverse environments are present. The highest accuracy is achieved by the NGCCs and the PLPs with an accuracy score of 0.850. Next, we perform an equivalent procedure and extract these features on the IMF basis functions of the correspondent dataset. Results are in table 7. Our proposed methodology relies on this new approach for which, instead of the raw speech, each IMF basis is passed one by one through an individual transformation of the selected features (i.e. MFCCs, LFCCs, BFCCs, etc.) to form adaptive features for the classification of real and synthetic voice. The standard practice of classification problems for this kind of setting is constructing a vector collecting all the coefficients for the feature of interest or multiple features and then carrying the learning procedure. Since voice is highly biometric, highly non-stationary and adverse environments might arise, such standard procedures tend to create noise in the classification tasks rather than provide discriminant information.

136844
Therefore, our idea is to partition the time-frequency plane through a non-stationary and non-linear decomposition method and quantify energy generated by the formant structure. Furthermore, depending on the targeted task, i.e. TD-SD-SV or TI-SI-SV, the discriminant areas might differ given the use of the same utterances or not or the presence of multiple speakers or not or the consideration of gender. We observe that all features (except for the LPCs and the PSRCCs) achieve higher accuracy scores on the IMFs, particularly on the highest bases as IMF1 or IMF2, suggesting higher formants of the female speaker voices in a TD-SD-SV environment should provide most of the discriminant power. The MFCCs and the IMFCCs gave the highest accuracy scores. Given these performances, their interpretability and their wide use within SV tasks, we selected the MFCCs to construct new features combined with the EMD. Hence, we focus on the individual speech features MFCCs extracted on the IMFs and further discuss them in the following sections. They will be used to construct the EMD-MFCC MKL. Before that, we firstly provide further evidence to show how the EMD-MFCCs better capture formant structures compared to standard MFCCs on the raw speech data.

FORMANT DETECTION FOR REAL AND SYNTHETIC VOICE

In Figure 7 we demonstrate the wide-band spectrograms, which are plotted to visualise the formant structure of a given speech signal. The four panels represent the same sentence for Speaker 1, Speaker 2, and the female synthetic voice and the male synthetic voice. Each spectrogram has been performed on a window of 1024 samples (corresponding approximately to 23 milliseconds), with an overlap of 128 samples, the same pre-emphasis factor and windowing applied for the MFCCs (0.97 and hamming window), a dynamic range of 50dB and frequency range set of 0-10 kHz so that five formants should be visible (one at each 1 kHz spaced carrier frequency). In [24] it is noted that the first five formants are the ones necessary for speaker verification. Black lines highlight the five detected formants over time in each sub-figure, which line up with the EMD decomposed IMFs after transformation to IFs. The top panel corresponds to the female speaker. The first four formants are within 0-5 kHz. This confirms that female speakers tend to have higher formant frequencies due to smaller vocal tracts (see [38]) and a higher fundamental frequency $F_0$ compared to males. The second panel shows five formants in the interval 0-5 kHz, typical for a male voice. Furthermore, a lower fundamental frequency generates a smaller interval between voice harmonics resulting in a strengthened formants definition. This shows that for male voice EMD decomposition versus female voice, the IMFs obtained will have energy concentration in different spectrum regions. Consequently, the resulting EMD-MFCC coefficients, if the Mel Cepstrum bases are kept constant in both cases, will have the coefficients for lower order IMFs being more influential than higher order IMFs. The opposite will be valid for the female voice. Note how the first two spectrograms show how human voices enunciate individual words much more than the last two spectrograms, where the separation between them seems dissipated. Furthermore, the formant structures referring to female voices (the first and the third plots) appear to behave much more alike than those characterising the male voices (the second and the fourth plots). This fact strongly depends on the synthetic voice generation algorithm, which will spread energy across a significant range of frequencies even by synthesising a male voice. Such a fact will result in a less challenging task for detecting synthetic and real male voices than the female case, and it justifies our choice to focus on the female case.

Next, to illustrate the EMD-MFCC method versus the classical MFCC on the raw speech, we selected a sentence randomly from the real female voice recordings, and we present the speech signal in the time domain and the spectrogram in Figure 8.

Then we plot the results of the $PM^m_k(h) = |\Gamma^m_k(h)| \cdot H(h, m)$ which are the PSD weighted Mel Cepstral bases for indexes $m \in \{1, \ldots, 12\}$ in Figure 9. We compare the classical situation in which one applies the MFCC directly to the speech signal (Figure 9, panel a) to the cases in which the MFCC are instead applied to each IMF individually, precisely the first three IMFs (Figure 9, panels b,c,d). In each case, we do this over the entire time interval of the recording, followed by a sequence of local MFCC applications on 200ms windows with no overlap. This demonstrates that the resulting MFCC summary features captured by $PM^m_k(h)$ applied to the raw data signal are not overly responsive, although, in adjacent 200ms
Since the IMF-MFCC features adapt to local non-stationarity better than the ST-DFT MFCC analysis, we can more responsively capture the energy variation in bands of the formants to a greater degree. In the subsequent classification of the biometric speech attack analysis, we will demonstrate that this leads to demonstrably better performance in our proposed method over state of the art methods.

In general, we observe individual features from the summary statistics of the IMF bases and, importantly, its improved responsivity and selectivity for formants based on the IMF-MFCC framework we propose.

In Figure 10, we explore the IMF-MFCC discriminatory potential of features between real and synthetic voice via application of the t-SNE projection method, see details in [69]. The details of the t-SNE technique and how we utilised this method for our case analysis are summarised in the Supplement Materials. The plots demonstrate the discriminatory power of the IMF-MFCC coefficient representations when applied on local windows of length 50ms, producing features vectors in dimension \(d = 1068\), after decimation for dimension reduction. One can see that there is evident potential for these IMF-MFCC features to have strong discriminatory power in all IMFs for the male case and in several IMFs for female ones. As expected, the female lower frequency IMF-MFCC features have less discriminatory power than the higher frequency signatures, and the IMF-MFCC captures this clearly in all sentences as discriminatory between the real and synthetic voice. This indicates that the IMF-MFCC should act very well as spectral signatures to capture an individual’s particular vocal tract structure and therefore have a solid performance to mitigate attacks from the synthetic voice.

**SVM-SPEECH FEATURE LIBRARY CONSTRUCTION: CLASSIFICATION PERFORMANCE FOR INDIVIDUAL SPEECH FEATURES**

Note that we focus on the female voice examples to present the results and provide similar results for the male voice in the supplementary appendix. In table 6, we see that applying MFCC to raw speech produces an accuracy of discrimination between real female voice and synthetic female attack spoofed voices, out of sample for the same sentences, which did not exceed 77.5%. This type of accuracy score is often not acceptable for real-world applications where sensitive private data is seeking to be accessed via voice biometrics.

The benchmark MFCCs on raw speech is further compared to two sets of features individually trained and tested. The first set corresponds to summary statistics obtained from the EMD applied to the raw speech signal. This produces the summary statistics of the IMF bases, the Instantaneous Frequency signals and the spline coefficients that characterize the IMF bases. We summarize these three signals using the summary statistics described in Table 1. These results are in Tables 1 and 2 in the Supplement Materials and demonstrate the out-of-sample classification results for dataset one for Speaker 1 versus synthetic voice attacker and Speaker 2 versus synthetic voice attacker.

We performed one SVM training and then out-of-sample testing per feature component, where, for instance, we took each sentence and then took each IMF. After, given each IMF, we extracted summary statistics and then ran the SVM training and out-of-sample testing for various kernel families. This allows building a library of individual features and their performance in the real vs synthetic voice discrimination over the voice recordings database. It forms the basis for the multiple kernel learning framework that ultimately creates our proposed EMD-MFCC multi-kernel classification solution. We bold all performances greater than 90% accuracy when presenting the results, which is a realistic minimum accuracy required for many real-world biometric applications. In general, we observe individual features from the summary statistics of the IMFs and IFs (for a range of kernel choices) outperforming the standard comparison of MFCC applied to raw speech in both in-sample out-of-sample analyses. This indicates that the approach we advocate for constructing IMF-MFCC rather than just MFCC on raw speech signals will outperform the current standard approach in this type of cyber mitigation ASV classification context.

The second set of features is obtained from an EMD applied to speech to get IMF bases, then the MFCCs are extracted from each IMF. This is our newly proposed methodology which utilized EMD-MFCC. Table 6 shows the out-of-sample results for the EMD-MFCCs of the female voice for dataset one, and table 3 in the Supplement Materials presents the correspondent results for the male voice. The results are...
FIGURE 9. This Figure shows four panels. By looking at panel (a), seven subplots can be found. The first and biggest subplot represents the $\text{PM}^*$ component of the MFCC decomposition presented in Eqn. 17 for one of the original speech signals of Speaker 1 (the female voice). We afterwards extract the same quantity over batches of $s(t)$, such as shown in the subfigures below such biggest plot. Panel (b), (c) and (d) take instead into account the correspondent $\text{PM}^*_1$, $\text{PM}^*_2$ and $\text{PM}^*_3$ components of the MFCC decomposition of $\gamma_1(t)$, $\gamma_2(t)$ and $\gamma_3(t)$, i.e. the first, the second and the third IMFs respectively of the original speech signal considered in panel (a). The time unit of the batches is in ms, and the frequency on the x-axis is in Hz. The y-axes of $\text{PM}^*_1$, $\text{PM}^*_2$ and $\text{PM}^*_3$ differ from the y-axis of $\text{PM}^*$ since the IMFs do not include the residual or tendency.

FIGURE 9. This Figure shows four panels. By looking at panel (a), seven subplots can be found. The first and biggest subplot represents the $\text{PM}^*$ component of the MFCC decomposition presented in Eqn. 17 for one of the original speech signals of Speaker 1 (the female voice). We afterwards extract the same quantity over batches of $s(t)$, such as shown in the subfigures below such biggest plot. Panel (b), (c) and (d) take instead into account the correspondent $\text{PM}^*_1$, $\text{PM}^*_2$ and $\text{PM}^*_3$ components of the MFCC decomposition of $\gamma_1(t)$, $\gamma_2(t)$ and $\gamma_3(t)$, i.e. the first, the second and the third IMFs respectively of the original speech signal considered in panel (a). The time unit of the batches is in ms, and the frequency on the x-axis is in Hz. The y-axes of $\text{PM}^*_1$, $\text{PM}^*_2$ and $\text{PM}^*_3$ differ from the y-axis of $\text{PM}^*$ since the IMFs do not include the residual or tendency.

Presented for the radial basis function kernel choice, and the remaining results for other kernel choices are similar in performance, so we omit them to reduce space. The way to interpret these results is as part of a stage of constructing a library of individual feature sets to pass to a multiple-kernel learning solution.
FIGURE 10. The panels represent the coefficient functions $\mathcal{M}_k(s)$ given in Eqn. 17 computed on a sliding window for one sentence of Speaker 1. Note that panel (a) refers to the original signal, and the correspondent quantity is denoted as $\mathcal{M}(s)$ with no sub-index. We split the sentence 200ms windows and calculated $\mathcal{M}(s)$ for every window. We then repeated the procedure on the IMFs basis of the same sentence of Speaker 1 and showed the results in the remaining panels obtaining $\mathcal{M}_1(s)$, $\mathcal{M}_2(s)$, $\mathcal{M}_3(s)$, $\mathcal{M}_K(s)$ and $\mathcal{M}_{K+1}(s)$. Remark that $K$ is the last IMF and, in this specific case, equals 14 and $K + 1$ corresponds to the residual. The different colours denote the associated window over which the extraction has occurred. Remark that the x-axes differ amongst the panels since the IMFs do not take into account the residual.

EMD-MFCC MULTI KERNEL LEARNING SVM PERFORMANCE

We now present our proposed solution by combining the selection of the best-performing features from the EMD along with the EMD-MFCC SVM feature libraries we have constructed for various kernel choices and individual features above described (note that the results for the individual features are within the Supplement Material). We achieved the combination through the Multi Kernel Learning (MKL) introduced in V-A. By being the best performing within the individual features studies, the EMD-MFCCs are selected in this task and demonstrated for the female case as the most challenging. Each of these chosen features (individually trained in previous experiments) will be combined according to eqn. 26. The procedure consists of selecting the best kernel amongst the best feature for each feature. As a consequence, our final combined kernel should be more representative of the classification problem. Table 8 displays results for Speaker 1 out-of-sample performance for dataset 1. Since we select the best performing EMD-MFCC features amongst several kernels, the header of table 8 is organised as follows: the top row shows which is the basis of interest, i.e. $\gamma_1(t')$, $\gamma_2(t')$, $\gamma_3(t')$, $\gamma_K(t')$ and $\gamma_{K+1}(t')$. The index following MFCC- gives this information. The second row highlights the best individually performing coefficient and, therefore, the one selected for the MKL formulation. The last row shows which kernel offers the best performance for that feature; for example, for column one of the table, for the first IMF $\gamma_1(t')$ (MFCC-1), the best performing coefficient was the 7-th one.
when a Laplace kernel was used. The rest of the columns can be interpreted equivalently. We then put the header referring to the weights $\eta_m$ for $m = 1, \ldots, 5$. Each row represents a new model and shows the weights $\eta_m$ defined in eqn. 26, which are associated with features given at the head of the table. When considered individually, they reflect their out-of-sample performances and, therefore, reflect which feature provides more significant discrimination. Thus, the rows provide the new models’ characterisation obtained through the combination rule given in eqn. 25 with related performance in the last column provided by the accuracy score. Note that performances are ordered according to the level of accuracy achieved.

Perfect discrimination is achieved when the MFCCs of $\gamma_1(t')$ or $\gamma_2(t')$ are included within the combined kernel. Such findings reinforce the initial analysis of the t-SNE that most of the discrimination between a real female voice and a synthetic female voice lies in the high-frequency MFCC coefficients of
TABLE 6. Out-of-sample results of the SVMs carried with the standard features used in ASV tasks applied to the raw data. The features description is given in table 5. Equivalent results for these features applied to the IMFs are provided in table 7. Note that each value corresponds to the accuracy achieved by the SVM carried with the coefficient given in the row of the feature given in the column.

| Coeff. number | MFCCs | LPCCs | BFCcs | GCFCcs | IMFCCs | LPCs | LPCCs | MSRCCs | NGCCcs | PLPs | PSRs | RPLPs |
|---------------|-------|-------|-------|--------|--------|------|-------|--------|--------|------|------|-------|
| 1             | 0.775 | 0.500 | 0.500 | 0.425 | 0.500 | 0.425 | 0.750 | 0.500  | 0.500  | 0.725 | 0.725 | 0.725 |
| 2             | 0.675 | 0.500 | 0.535 | 0.775 | 0.725 | 0.725 | 0.750 | 0.535  | 0.700  | 0.535 | 0.535 | 0.535 |
| 3             | 0.500 | 0.500 | 0.500 | 0.775 | 0.400 | 0.725 | 0.725 | 0.500  | 0.725  | 0.650 | 0.725 | 0.725 |
| 4             | 0.475 | 0.500 | 0.700 | 0.525 | 0.075 | 0.725 | 0.700 | 0.475  | 0.850  | 0.850 | 0.375 | 0.450 |
| 5             | 0.475 | 0.500 | 0.650 | 0.475 | 0.775 | 0.425 | 0.775 | 0.500  | 0.775  | 0.450 | 0.450 | 0.450 |
| 6             | 0.650 | 0.500 | 0.700 | 0.775 | 0.725 | 0.325 | 0.700 | 0.450  | 0.700  | 0.435 | 0.435 | 0.435 |
| 7             | 0.150 | 0.475 | 0.700 | 0.725 | 0.725 | 0.325 | 0.700 | 0.500  | 0.750  | 0.700 | 0.525 | 0.700 |
| 8             | 0.375 | 0.500 | 0.700 | 0.525 | 0.725 | 0.725 | 0.675 | 0.425  | 0.800  | 0.700 | 0.475 | 0.675 |
| 9             | 0.375 | 0.500 | 0.775 | 0.725 | 0.725 | 0.725 | 0.225 | 0.500  | 0.750  | 0.460 | 0.650 | 0.700 |
| 10            | 0.675 | 0.550 | 0.825 | 0.800 | 0.700 | 0.725 | 0.750 | 0.500  | 0.750  | 0.450 | 0.475 | 0.475 |
| 11            | 0.770 | 0.500 | 0.700 | 0.800 | 0.775 | 0.725 | 0.750 | 0.500  | 0.775  | 0.500 | 0.225 | 0.500 |
| 12            | 0.775 | 0.500 | 0.700 | 0.725 | 0.750 | 0.725 | 0.725 | 0.500  | 0.775  | 0.550 | 0.550 | 0.500 |

the first IMF. Indeed, the selected coefficients for this case corresponds to the 7-th. Different combinations have been tried, and similar excellent performance was observed. These results are far higher than those of the current state-of-the-art reference of MFCC applied to raw speech when also placed in an MKL-SVM framework. This demonstrates the superior performance of the IMF-MFCC feature class when combined with an MKL-SVM classifier framework. Note that each EMD-MFCC-MKL table will follow the structure described in this section. Note that the individual performances related to the MFCCs on the raw speech signals are not presented for other datasets.

HARVARD PHONETICALLY BALANCED SENTENCES EXAMPLE: REAL SPEECH VS SYNTHETIC SPEECH CLASSIFICATION
As performed for dataset one, a similar analysis was confirmed by dataset two on the gold standard speech data set given by the Harvard phonetically balanced sentences. Table 9 shows results related to Speaker 1, hence the female discrimination case study (as per the above study one). We provide a summary of the EMD-MFCC features only and MKL-SVM classifier compared to the MFCC on raw speech in an MKL-SVM classifier framework. Note that each EMD-MFCC-MKL table will follow the structure described in this section. Note that the individual performances related to the MFCCs on the raw speech signals are not presented for other datasets.

B. EXPERIMENT TWO: OTHER TTS ALGORITHMS
In this subsection, we replicate the EMD-MFCC-MKL conducted in experiment one by taking into account different Text-To-Speech (TTS) algorithms, presented in Table 10. Note that we replicate the experiment for the female voice only but both dataset one and dataset two. The first TTS algorithm corresponds to the interface of the Google-Text-to-Speech API provided by the Python library gTTS. It relies on WaveNet [62] and hence uses a Deep Learning procedure. It offers 120 languages and dialects (see https://cloud.google.com/speech-to-text/docs/languages). The second TTS algorithm corresponds to Espeak (online at http://espeak.sourceforge.net/) that instead employs a formant synthesis procedure. It also provides several languages (the complete list is given online). Afterwards, we use the Python library Pyttsx, a cross-platform text-to-speech wrapper providing access to different TTS tools. Amongst others, we select the Microsoft Speech Engine SAPI5 (online at https://docs.microsoft.com/en-us/previous-versions/windows/desktop/ms723627(v=vs.85)), making use of a concatenative algorithm. The last TTS algorithm is the IBM Watson TTS (whose documentation can be found online at https://cloud.ibm.com/docs/text-to-speech), which also provides access to its API through a Python interface and relies on neural voice technologies, hence making use of Deep Neural Network (DNN). The TTS service is in the IBM Watson Cloud and supports a large number of languages, from which we selected the option of UK English.

As in experiment one, we firstly carry out individual feature SVMs, hence one for each mel-frequency cepstral coefficient of the obtained IMFs. Results concerning these SVMs are provided in the Supplement Materials in tables 4 and 5 for dataset one and tables 6 and 7 for dataset two. We then selected the best performing cepstral coefficients per IMF basis function and then carried out the EMD-MFCC-MKL procedure as presented in experiment one. Results for the IBM TTS algorithm are provided in tables 11 and 12 for dataset one and dataset two, respectively. Results for the remaining algorithms are in Supplement Materials in tables 8, 9, 10 and 11. As in the previous experiments, the best performing MFCCs for the first three IMFs are high-frequency ones confirming our initial claim that most of the discrimination power for female voices should come from these regions of the time-frequency plane.
TABLE 7. Out-of-sample results of the SVMs carried with the standard features used in ASV tasks applied to the IMFs. The features description is given in Table 5. Results for these features applied to the raw data are provided in Table 6. Note that each value corresponds to the accuracy achieved by the SVM carried with the coefficient given in the row of the IMF basis in the column referring to the feature provided.

C. EXPERIMENT THREE: APPLICATION ON THE ASVspoof 2019 CHALLENGE DATASET

One of the biggest problems affecting AV studies is the comparison of various techniques evaluated over different datasets. As a result, since 2015, the research community [70]–[73] has started to release evaluation databases as SAS, ASVspoof 2015, ASVspoof 2017, ASVspoof 2019 challenge, AVspoof, RedDots Replayed databases. Consistently with these purposes, we investigate the new technique combining EMD-MFCCs by employing the ASVspoof 2019 challenge database [55]. Table 13 (also at https://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf) describes the structure of such a dataset. It subdivides into two different scenarios: logical access (LA) and physical access (PA). Hence, replay spoofing attacks are VOLUME 9, 2021

V. CONCLUSION

81851
FIGURE 12. Pros and Cons of TTS algorithms.

**CONS:**
- (1) Time-Consuming Systems requiring large databases and hard-coding the combination to form the words.
- (2) The resulting speech may sound less natural and emotionless since it is hard to get the audio recordings of all possible words spoken in all possible combinations of emotions, prosody, etc.

**PROS:**
- (1) High quality of audio in terms of intelligibility.
- (2) Possibility to preserve the original actor’s voice.

**FIGURE 13.** Extracted spoofing attacks for experiment three from the ASVSpoof 2019 challenge database from the Logical Access set.

**LOGICAL ACCESS - SPOOFING ATTACKS**

| Subset | # of Attacks | Known | Unknown | Type of Attack |
|--------|--------------|-------|---------|----------------|
| Training | 6            | Known | 0       | VC and TTS     |
| Transient | 6            | Known | 0       | VC and TTS     |
| Evaluation | 10          | Known | 11      | VC, TTS, Hybrid |

**EXTRACTED SPOOFING ATTACKS**

| Subset | # of Attacks | Known | Unknown | Type of Attack | Algorithm          |
|--------|--------------|-------|---------|----------------|--------------------|
| Training | 5            | Known | 0       | TTS            | Deep learning, Concatenative |
| Development | 2         | Known | 0       | TTS            | Deep learning, Concatenative |

recordings of bonafide speech assumed to be captured and then represented to the microphone of an ASV system using a replay device. In this work, we only take into account the Logical Access scenario to target the TTS algorithms used within this database. The LA database contains bonafide speech and spoofed speech data obtained using 17 different TTS and VC systems. Figure 13 shows its spoofing attacks structure and the ones we extracted for our experiments. Note that data for the training of TTS and VC systems partly comes from the VCKT database (online at http://dx.doi.org/10.7488/ds/1994), but there is no overlap with the data contained in the 2019 database. Among the 17 spoofing voice generation systems, 6 are known attacks, while 11 are unknown. The training and development sets contain known attacks only, while the evaluation set contains 2 known and 11 unknown spoofing attacks. Regarding the 6 known attacks, there are 2 VC systems and 4 TTS systems. Particularly, VC systems use neural-network-based and spectral-filtering-based approaches [74]. TTS use either waveform concatenation or neural-network-based speech synthesis using a conventional source-filter vocoder [75] or a WaveNet-based vocoder [62]. We extract three of the TTS spoofing attacks for the training and development sets.

The generation algorithms for the spoof voices fall into either Deep Learning or Concatenative types and can be compared to our results (see table 10). In particular, the chosen algorithms are decoded in the LA protocol of the ASVSpoof 2019 challenge as “A01”, “A02” and “A04”, respectively. The TTS “A01” algorithm is obtained with a neural waveform model, while the TTS “A02” algorithm is generated through a vocoder. Lastly, the TTS “A04” algorithm is a waveform concatenation. There are 8 male and
TABLE 8. Multi kernel learning SVMs results of the synthetic voice generated with TTS T1 versus Speaker 1 for dataset 1. We select the best features according to their performances when individually tested (i.e. through the out-of-sample accuracy). The first line indicates the considered features, which is always an IMF-MFCC. The IMF indices are given in each MFCC component as \( -1, -2, -3, -K, -K+1 \). The second line refers to the coefficient number, and the third line to the selected kernel for that feature. The table represents a model selection comparison in which each row corresponds to a different MKL model combining different sets of features. The numbers in each row refer to the \( \eta_m \) weights as expressed in Eqn. 26. The highlighted accuracy scores correspond to those combinations of features and kernel models greater than 90%. The first portion of the table demonstrates the EMD-MFCC-MKL solutions, while the second portion is the state-of-the-art reference of the classical MFCC-MKL.

| Experiment 1 | Dataset 1 |
|--------------|-----------|
| OFS EMD-MFCC-MKL - Speaker 1 versus Female Synthetic Voice TTS T1 | |
| MFCC-1 | MFCC-2 | MFCC-3 | MFCC-K | MFCC-K+1 | 7th coeff. | 5th coeff. | 10th coeff. | 5th coeff. | Accuracy |
| Laplace | RBF | RBF | RBF | RBF | Linear |
| \( \eta_1 \) | \( \eta_2 \) | \( \eta_3 \) | \( \eta_4 \) | \( \eta_5 \) | |
| 0.295 | 0.295 | 0.222 | 0.188 | -- | 1.000 |
| 0.364 | 0.364 | 0.272 | -- | -- | 1.000 |
| 0.300 | 0.260 | 0.240 | -- | -- | 1.000 |
| 0.422 | -- | 0.311 | 0.267 | -- | 1.000 |
| -- | 0.416 | 0.314 | 0.269 | -- | 1.000 |
| 0.500 | 0.500 | -- | -- | -- | 1.000 |
| 0.572 | -- | -- | -- | -- | 1.000 |
| 0.610 | -- | -- | 0.390 | -- | 1.000 |
| -- | 0.611 | -- | 0.389 | -- | 0.990 |
| 0.246 | 0.246 | 0.184 | 0.157 | 0.167 | 0.900 |
| 0.292 | 0.292 | 0.216 | -- | 0.200 | 0.990 |
| -- | 0.328 | 0.246 | 0.208 | 0.218 | 0.900 |
| 0.418 | 0.048 | -- | 0.262 | 0.272 | 0.890 |
| 0.325 | -- | -- | 0.242 | 0.214 | 0.890 |
| 0.374 | 0.374 | -- | -- | -- | 0.890 |
| 0.415 | -- | 0.309 | -- | 0.276 | 0.890 |
| -- | 0.362 | 0.309 | 0.329 | 0.890 |
| -- | 0.436 | -- | 0.277 | 0.287 | 0.890 |
| -- | 0.413 | 0.310 | -- | 0.277 | 0.890 |
| -- | 0.573 | 0.427 | -- | -- | 0.890 |
| -- | 0.602 | -- | 0.529 | 0.471 | 0.890 |
| 0.435 | -- | -- | 0.274 | 0.291 | 0.880 |
| 0.600 | -- | -- | -- | 0.400 | 0.880 |
| -- | -- | -- | 0.486 | 0.514 | 0.870 |

TABLE 9. Multi kernel learning SVMs results of the synthetic voice generated with TTS T1 versus Speaker 1 for dataset 2. We select the best features according to their performances when individually tested (i.e. through the out-of-sample accuracy). The first line indicates the considered features, which is always an IMF-MFCC. The IMF indices are given in each MFCC component as \( -1, -2, -3, -K, -K+1 \). The second line refers to the coefficient number, and the third line to the selected kernel for that feature. The table represents a model selection comparison in which each row corresponds to a different MKL model combining different sets of features. The numbers in each row refer to the \( \eta_m \) weights as expressed in Eqn. 26. The highlighted accuracy scores correspond to those combinations of features and kernel models greater than 90%. The first portion of the table demonstrates the EMD-MFCC-MKL solutions, while the second portion is the state-of-the-art reference of the classical MFCC-MKL.

| Experiment 1 | Dataset 2 |
|--------------|-----------|
| OFS EMD-MFCC-MKL - Speaker 1 versus Female Synthetic Voice TTS T1 | |
| MFCC-1 | MFCC-2 | MFCC-3 | MFCC-K | MFCC-K+1 | 8th coeff. | 7th coeff. | 10th coeff. | 5th coeff. | Accuracy |
| RBF | RBF | RBF | RBF | RBF | Linear |
| \( \eta_1 \) | \( \eta_2 \) | \( \eta_3 \) | \( \eta_4 \) | \( \eta_5 \) | |
| 0.170 | 0.170 | 0.235 | 0.268 | 0.157 | 0.944 |
| 0.201 | 0.201 | 0.274 | 0.324 | -- | 0.944 |
| 0.233 | 0.241 | 0.318 | -- | 0.208 | 0.923 |
| -- | 0.208 | 0.280 | 0.328 | 0.184 | 0.923 |
| 0.397 | 0.409 | 0.194 | -- | -- | 0.923 |
| -- | 0.253 | 0.341 | 0.406 | -- | 0.923 |
| -- | 0.308 | 0.416 | -- | 0.276 | 0.999 |
| 0.218 | 0.226 | -- | 0.352 | 0.204 | 0.902 |
| 0.203 | -- | 0.281 | 0.328 | 0.188 | 0.895 |
| 0.275 | 0.283 | -- | 0.442 | -- | 0.861 |
| -- | 0.289 | -- | 0.454 | 0.257 | 0.861 |
| 0.336 | 0.355 | -- | -- | 0.309 | 0.861 |
| 0.491 | 0.509 | -- | -- | -- | 0.861 |
| -- | 0.426 | 0.574 | -- | -- | 0.861 |
| -- | 0.388 | -- | 0.612 | -- | 0.861 |
| -- | 0.531 | -- | -- | 0.469 | 0.861 |
| 0.230 | -- | 0.348 | 0.402 | -- | 0.833 |
| 0.301 | -- | 0.420 | -- | 0.279 | 0.833 |
| -- | -- | 0.353 | 0.413 | 0.234 | 0.833 |
| 0.285 | -- | -- | 0.460 | 0.255 | 0.500 |
| 0.418 | -- | 0.582 | -- | -- | 0.500 |
| 0.382 | -- | -- | 0.618 | -- | 0.500 |
| 0.521 | -- | -- | 0.479 | -- | 0.500 |
| -- | -- | 0.603 | 0.537 | -- | 0.500 |
| -- | -- | -- | 0.638 | 0.362 | 0.500 |

12 female speakers for bonafide speech utterances and the selected TTS algorithms in the training set. However, there are 12 female and 8 male voices for the bonafide speech in the development set, but 6 female and 4 male voices for the selected synthetic ones. Note that the speakers differ between the training and the development sets, and the utterances differ amongst the speakers. Hence, a text-independent and speaker-independent scenario is the one of interest in this experiment (TI-SI-SV). Furthermore, the number of utterances per speaker and type of speech (i.e. bonafide or synthetic) differ. Therefore, the selected subset is unbalanced in terms of the number of utterances in the bonafide or natural set versus the spoof groups. This results from a different number of speakers’ utterances (this information is not evident in the proposed tables). As a result, we balance both the training and development subsets. For the training set, we decide to select the minimum number of utterances available for one speaker and then randomly select the same number from OFS Raw Data MFCC-MKL - Speaker 1 versus Female Synthetic Voice TTS T1 | |
| MFCC-1 | MFCC-2 | MFCC-3 | MFCC-K | MFCC-K+1 | 8th coeff. | 7th coeff. | 10th coeff. | 5th coeff. | Accuracy |
| RBF | RBF | RBF | RBF | RBF | Linear |
| \( \eta_1 \) | \( \eta_2 \) | \( \eta_3 \) | \( \eta_4 \) | \( \eta_5 \) | |
| 0.202 | 0.199 | 0.199 | 0.199 | 0.199 | 0.712 |

VOLUME 9, 2021
TABLE 10. Table describing the Text-To-Speech (TTS) Tools employed in experiment two for comparisons of different Speech Synthesis algorithms producing different synthetic voices and generating different types of attacks. Note that speech generated through TTS T1, corresponding to the online TTS, was obtained from http://www.fromtexttospeech.com/.

| Experiment 2 |
|--------------|
| # | TTS Tool | Algorithm | Gender | Accent |
|---|----------|-----------|--------|--------|
| T1 | Online TTS | Stat. Par. | Male | British En., Harry |
| T1 | Online TTS | Stat. Par. | Female | British En., Emma |
| T2 | Google TTS | Deep Learning | Female | British En., m-OB-Wavenet-A |
| T3 | IBM Watson TTS | Deep Learning | British En., Hazel |
| T4 | IBM Watson TTS | Deep Learning | British En., Charlotte |
| T5 | SAPS | Conspicuous | Female | British En., Mary |

TABLE 11. Multi kernel learning SVMs results of the synthetic voice generated with the IBM TTS algorithm versus Speaker 1 for dataset 1. We select the best features according to their performances when individually tested (i.e. through the out-of-sample accuracy). The first line indicates the considered features, which is always an IMF-MFCC. The IMF indices are given in each MFCC component as $-1, -2, -3, \ldots, -K, +1$. The second line refers to the coefficient number, and the third line to the selected kernel for that feature. The table represents a model selection comparison in which each row corresponds to a different MKL model combining different sets of features. The numbers in each row refer to the $\eta_m$ weights as expressed in Eqn. 26. The highlighted accuracy scores correspond to those combinations of features and kernel models greater than 90%. The first portion of the table demonstrates the EMD-MFCC-MKL solutions, while the second portion is the state-of-the-art reference of the classical MFCC-MKL.

| Dataset 1 |
|----------|
| OFS EMD-MFCC-MKL - Speaker 1 vs Female Synthetic Voice IBM |
| MFCC-1 | MFCC-2 | MFCC-3 | MFCC-K | MFCC-K +1 |
| 8th coeff. | 9th coeff. | 9th coeff. | 11th coeff. | Accuracy |
| RBF | RBF | RBF | RBF | RBF |
| $\eta_1$ | $\eta_2$ | $\eta_3$ | $\eta_4$ | $\eta_5$ |
| 0.260 | 0.250 | 0.220 | 0.135 | 0.159 | 1.000 |
| 0.305 | 0.297 | 0.288 | 0.170 | 1.000 |
| – | 0.609 | 0.397 | 0.194 | 1.000 |
| – | 0.331 | 0.291 | 0.189 | 1.000 |
| 0.341 | 0.281 | 0.189 | 0.189 | 0.599 |
| 0.361 | 0.351 | 0.288 | – | 0.599 |
| 0.422 | 0.410 | – | 0.168 | 0.599 |
| 0.412 | 0.401 | – | 0.187 | 0.597 |
| – | 0.000 | 0.389 | – | 0.131 |
| 0.301 | 0.292 | 0.284 | 0.123 | 0.596 |
| 0.424 | – | 0.400 | 0.176 | 0.596 |
| 0.433 | – | 0.409 | 0.158 | 0.595 |
| 0.436 | – | 0.282 | 0.282 | 0.593 |
| 0.537 | 0.463 | – | – | 0.592 |
| 0.556 | – | 0.444 | – | 0.591 |
| 0.349 | – | 0.329 | 0.161 | 0.590 |
| 0.605 | – | – | 0.395 | – |
| – | 0.434 | 0.273 | 0.273 | 0.589 |
| – | 0.436 | 0.282 | 0.282 | 0.589 |
| – | 0.620 | 0.380 | 0.380 | 0.796 |
| 0.623 | – | 0.377 | – | 0.795 |
| 0.636 | – | 0.364 | – | 0.794 |
| – | 0.669 | 0.331 | – | 0.793 |
| 0.520 | 0.480 | – | – | 0.792 |
| 0.686 | – | – | 0.314 | 0.789 |
| MFCC | MFCC | MFCC | 8th coeff. | Accuracy |
| RBF | RBF | RBF | RBF | RBF |
| $\eta_1$ | $\eta_2$ | $\eta_3$ | $\eta_4$ | $\eta_5$ |
| 0.202 | 0.200 | 0.200 | 0.198 | 0.675 |

TABLE 12. Multi kernel learning SVMs results of the synthetic voice generated with the IBM TTS algorithm versus Speaker 1 for dataset 2. We select the best features according to their performances when individually tested (i.e. through the out-of-sample accuracy). The first line indicates the considered features, which is always an IMF-MFCC. The IMF indices are given in each MFCC component as $-1, -2, -3, \ldots, -K, +1$. The second line refers to the coefficient number, and the third line to the selected kernel for that feature. The table represents a model selection comparison in which each row corresponds to a different MKL model combining different sets of features. The numbers in each row refer to the $\eta_m$ weights as expressed in Eqn. 26. The highlighted accuracy scores correspond to those combinations of features and kernel models greater than 90%. The first portion of the table demonstrates the EMD-MFCC-MKL solutions, while the second portion is the state-of-the-art reference of the classical MFCC-MKL.

| Dataset 2 |
|----------|
| OFS EMD-MFCC-MKL - Speaker 1 vs Female Synthetic Voice IBM |
| MFCC-1 | MFCC-2 | MFCC-3 | MFCC-K | MFCC-K +1 |
| 8th coeff. | 9th coeff. | 9th coeff. | 6th coeff. | 1st coeff. | Accuracy |
| RBF | RBF | RBF | RBF | RBF |
| $\eta_1$ | $\eta_2$ | $\eta_3$ | $\eta_4$ | $\eta_5$ |
| 0.239 | 0.230 | 0.233 | 0.101 | 0.197 | 1.000 |
| 0.275 | 0.265 | 0.269 | 0.191 | – | 1.000 |
| 0.273 | 0.262 | 0.267 | – | 1.000 |
| 0.377 | 0.309 | 0.314 | – | 1.000 |
| 0.380 | 0.365 | 0.255 | – | 0.999 |
| 0.412 | 0.302 | 0.132 | 0.154 | – |
| 0.397 | 0.312 | 0.134 | 0.157 | – |
| 0.455 | 0.361 | – | 0.184 | 0.998 |
| 0.401 | – | 0.312 | 0.135 | 0.157 |
| 0.487 | – | 0.379 | 0.164 | 0.996 |
| 0.437 | – | 0.375 | – | 0.979 |
| – | 0.456 | – | – | 0.899 |
| 0.573 | – | 0.379 | 0.163 | 0.898 |
| 0.537 | – | 0.197 | 0.230 | 0.896 |
| 0.438 | 0.374 | – | 0.188 | 0.899 |
| 0.324 | – | 0.476 | – | 0.879 |
| 0.571 | – | 0.198 | 0.231 | 0.868 |
| 0.733 | – | 0.267 | – | 0.868 |
| – | 0.525 | 0.475 | – | 0.799 |
| – | 0.609 | – | 0.391 | 0.799 |
| – | 0.695 | – | – | 0.305 |
| – | – | 0.591 | 0.203 | 0.236 |
| – | 0.642 | 0.358 | – | 0.768 |
| – | 0.639 | – | 0.371 | 0.779 |
| – | – | 0.544 | 0.456 | 0.768 |
| 0.623 | – | – | 0.377 | 0.706 |

TABLE 13. Summary of the ASVspoof 2019 challenge database as highlighted at https://www.asvspoof.org/index2019.html.

| Experiment 3 |
|--------------|
| ASVspoof 2019 Challenge database |
| # of Speakers | # of Utterances |
| Logical Access | Physical Access |
| Subset | Male | Female | Natural | Spof | Natural | Spof |
| Training | 8 | 12 | 2540 | 22800 | 5400 | 48600 |
| Development | 8 | 12 | 2548 | 22296 | 5400 | 24300 |
| Evaluation | – | 71747 | – | 137457 | – | – |

each of the other available speakers in every group (bonafide and spoofed). This corresponds to 127 utterances for every speaker, 2,540 utterances for every group (i.e. natural, A01, A02, A04), with a total of 10,160 utterances. Regarding the development subset, we first balanced the number of speakers within each gender and randomly selected the minimum...
TABLE 14. Summary of the extracted database from the ASVspoof 2019 challenge database to conduct our experiment three. Note that we selected two subsets, i.e. the training and the development. Furthermore, for the spoofed speech, we considered three of the TTS voices only. Note that the datasets is balanced in terms of number of utterances per speaker. We make use of the training set to train our SVMs proposed models and the development set for the testing.

| Subset     | Male | Female | Natural A01 | A02 | A04 |
|------------|------|--------|-------------|-----|-----|
| Training   | 8    | 12     | 2540        | 2540| 2540|
| Development| 4    | 4      | 616         | 616 | 616 |

TABLE 15. Multi kernel learning SVMs results of the female case versus the synthetic voice generated with the A0 TTS algorithm of the ASVspoof challenge dataset. We select the best features according to their performances when individually tested (i.e. through the out-of-sample accuracy).

| Experiment 3 | Dataset 3 |
|--------------|-----------|
| # of Speakers | # of Utterances |
| Subset        | Male | Female | Natural A01 | A02 | A04 |
| Training      | 8    | 12     | 2540        | 2540| 2540|
| Development   | 4    | 4      | 616         | 616 | 616 |

TABLE 16. Multi kernel learning SVMS results of the female case versus the synthetic voice generated with the A02 TTS algorithm of the ASVspoof challenge dataset. We select the best features according to their performances when individually tested (i.e. through the out-of-sample accuracy).

| Experiment 3 | Dataset 3 |
|--------------|-----------|
| # of Speakers | # of Utterances |
| Subset        | Male | Female | Natural A01 | A02 | A04 |
| Training      | 8    | 12     | 2540        | 2540| 2540|
| Development   | 4    | 4      | 616         | 616 | 616 |

number between the two, giving 4 male and 4 female voices in each group (bonafide, A01, A02 and A04). Furthermore, we applied the same procedure followed for the training set and randomly selected the minimum number of utterances available per speaker within each group corresponding to 77. Therefore, each group will have 616 utterances leading to 2464 utterances for the development set. For our experiments, we used the training set to train the individual features required to develop the EMD-MFCC-MKL and the development set for testing such a procedure with the added trait of gender, hence dividing the utterances according to it. Table 14 provides a summary of such a dataset. Each utterance speech recording duration was approximately 1sec to 3sec maximum sampled at 16kHz producing between 25k and 150k samples per spoken utterance. The start and end of each sample were trimmed to remove any non-speech segments
and decimated to a set of 40k total samples. The procedure concerning the EMD extraction followed the same applied to the other datasets, i.e. each set of 40k samples for one sentence was then windowed into non-overlapping collections of 5,000 samples and passed to the EMD sifting procedure. Then, for each IMF, we extracted M = 12 cepstral coefficients similarly to experiments one and two. We carried one individual SVM per coefficient per IMF for the female and male cases by considering the three different TTS algorithms. Results of the individual features are provided in the Supplement Materials in tables 14 and 15. For both genders, better performances are achieved by the MFCCs of the second or the third IMF basis function detecting lower speech formants and the fundamental frequency. In this context, multiple speakers are trained together through a unique model, and, particularly at high-frequencies of female voices, the non-stationarity of each speaker might be strongly biometric, resulting in out-of-sample accuracy levels of 70% for high cepstral coefficients of IMF1. What is instead detected more efficiently in a TI-SI-SV environment are lower formants and the fundamental frequency depicted by lower cepstral coefficients of IMF3. Therefore, the EMD-MFCCs provide interpretable high-performing features for this kind of speaker verification system. The following step corresponds to the EMD-MFCC-MKL analysis. Results for the female case versus the A01 and A02 TTS algorithms and the male case versus the same TTS algorithms are provided in tables 15, 16, 17 and 18. The other results considering the TTS algorithms A04 are in the Supplement Materials in tables 14 and 15. The MKL performances reinforce the findings related to the individual SVMs. In both female and male SVMs, highest accuracy levels (> 90%) are shown when the cepstral coefficients of IMF2 and IMF3 have been included in the MKL model. Furthermore, the male EMD-MFCC-MKL performances appear overall higher than the female ones; most of the formants lie, in male voices, at the lower frequency bandwidths and, compared to female
formants, present in general lower non-stationarity levels. Hence, better performances are achieved if low cepstral coefficients of IMF2 and IMF3 are considered. Furthermore, the EMD-MFCC-MKL framework provides a higher level of accuracy in every case compared to the individually trained EMD-MFCC. Indeed, in the latter case, no feature achieves an accuracy level greater than 90% (these are in tables 12 and 13 of the Supplement Materials). This strongly supports our proposed methodology. Regarding the TTS algorithms, A04, hence the concatenative approach, represents a more challenging spoofing attack than A01 and A02.

VIII. DISCUSSION AND CONCLUSION
A new speech biometric cyber-attack mitigation framework was developed in the class of ASV system. This allowed addressing the challenge of the classification of synthetic and real voices. Such a biometric security task needs to account for three main factors: firstly, speech is highly non-stationary and, therefore, methods that can depict such a property are required. Secondly, the fundamental characteristic of a speech signal is its formants structure. Since each individual has distinct vocal tracts, observing formants structure is the keystone in speech applications. Furthermore, measuring energy concentration around such frequencies should provide the discriminatory power required to differentiate spoofed and bonafide voices. Thirdly, the speech scenario considered provides different settings affecting the interpretation of the identified discrimination power. Hence, the flexibility of the classification technique in this respect is highly required. The method should be adaptive and interpretable, hence dependent on the given speech dataset but relying on a robust technique whose interpretation can be derived according to the scenario of interest (i.e. TD-SD-SV, TD-SI-SV, etc.).

Our proposed solution is achieved by building upon existing methodologies and adapting them to work with non-stationary signals more effectively. In this way, more robust features reducing sensitivity and enhancing performance in attack mitigation are achieved. Our robust method for speech synthesis spoofing attacks combines EMD and MFCCs with a multi-kernel learning SVM classifier framework. The new formulated feature libraries called EMD-MFCCs are explored and compared in various real data studies of different complexities. We demonstrate that, since the IMF's separate frequency bands of the original signals, the employment of the MFCCs relying on the mel-filter allows us to observe how frequency formants are concentrated in each IMF. The out-of-sample analysis offers better performances than the current state-of-the-art MFCC based solutions when applied directly to speech signals. We note that the current methodology of MFCC features applied directly to speech and utilised in a multi-kernel learning SVM could not achieve the minimum required standard for classification of 90% typical of biometric security. The new proposed methodology had many instances of out-of-sample performance with accuracies well above this threshold for all experiments taken into account.

The standard practice in these settings is to consider the MFCCs applied to the raw data and then construct a feature vector containing the entire set of coefficients. In this regard, we claim that the discrimination power identified by the classifier would be reduced and polluted by the different frequency bandwidths, and hence the different formants captured within a unique feature representation. The time-frequency plane must be partitioned with an a posteriori technique since the location of the formants is strictly individual-related and cannot be known a priori. Once this step is achieved, a parsimonious model trained with the computationally efficient classifier SVM-MKL is proposed. At this stage, we highlight that the “new-state-of-art” methods for speech classification tasks highly rely on DNNs. This class of methodologies requires a massive amount of data and high computational capacity due to the large volume of training required. The posed objective for a DNN applied in ASV settings, or equivalently in Automatic Speaker Recognition framework, is to learn individual or multiple speakers formants structure (depending on the selected speech scenarios) by training many layers of perceptrons. This procedure is replaced with the proposed methodology through a functional characterisation of the EMD and its basis functions. Therefore, rather than learning the formants through piece-wise functions using DNN complex layer structures, we extract them through the EMD and construct a simpler classifier. We propose a sparse architecture that replaces the DNN learning the formants with an EMD basis representation requiring far fewer parameters and can be applied to small and large datasets. It is computationally very efficient and, through an MKL ensemble method, achieves high accuracy levels in performances similar to the ones often achieved by the DNN when combined.

From a speech scenario perspective, text-dependent, speaker-dependent and text-independent, speaker-independent speaker verification systems have been tested. The proposed EMD-MFCC-MKL performed better than the standard benchmark features applied to the raw speech data in both cases. Furthermore, the created features have proven to produce interpretable machine learning solutions that provide flexibility for the targeted system. Several Text-To-Speech algorithms have been considered for the spoofing attacks in both the proposed scenarios and the studied features capture the synthetic voice better than standard ones. In the case of TI-SI-SV, the concatenative TTS algorithm appears to be the most difficult to capture in both female and male cases.

The proposed feature libraries correspond to are not overly engineered with excessive parametrisations. We showed that the EMD-MFCCs features offer the advantage of more reliable and robust MKL-SVM classifiers. As a result, they can be generalised in different non-stationary and noisy environments. This is particularly important in real-world situations usually associated with speech biometric access ASV technologies where a speaker may be providing a recording of speech through a non-ideal background noise mobile
environment. Hence, the signal transmission will not be subject to distortions, and the receiving device would then process reliable speech features to determine if access should be granted to sensitive data.

REFERENCES

[1] M. Y. Faisal and S. Suyanto, “SpecAugment impact on automatic speaker verification system,” in Proc. Int. Seminar Res. Inf. Technol. Intell. Syst. (ISRTI), Dec. 2019, pp. 305–308.

[2] K. Salah and S. Halim, “Kernel function and dimensionality reduction effects on speaker verification system,” in Proc. Int. Conf. Electr. Eng. (ICEE), Feb. 2020, pp. 1–4.

[3] K. J. Chakravishilvi and A. I. Prangshivil, “Verification in biometric systems: Problems and modern methods of their solution,” J. Appl. Statist., vol. 45, no. 1, pp. 43–62, Jan. 2018.

[4] N. E. Huang, Z. Shen, S. R. Long, C. Wu, H. H. Shi, Q. Zheng, N.-C. Yen, C. C. Tung, and H. L. Liu, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,” Proc. Roy. Soc. London A Math, Phys. Eng. Sci., vol. 454, no. 1971, pp. 903–995, Mar. 1998.

[5] R. Sharma, L. Vignol, G. Schlottmuer, M. A. Colomina, H. L. Rufiner, and S. R. M. Prasanna, “Empirical mode decomposition for adaptive AM-FM analysis of speech: A review,” Speech Commun., vol. 88, pp. 39–64, Apr. 2017.

[6] K.-H. Wu and C.-P. Chen, “Empirical mode decomposition for noise-robust automatic speech recognition,” in Proc. 11th Ann. Conf. Int. Conf. Spoken Lang. Proc. (ICSLP), Sep. 2009, pp. 3074–2077.

[7] V. Sethu, E. Ambikairajah, and J. Epps, “Empirical mode decomposition based weighted frequency feature for speech-based emotion classification,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., Mar. 2008, pp. 5017–5020.

[8] G. Schlottmuer, M. E. Torres, and H. L. Rufiner, “A new algorithm for instantaneous F0 speech extraction based on ensemble empirical mode decomposition,” in Proc. 17th Eur. Signal Process. Conf. (EUSIPCO), Sep. 2009, pp. 3247–2351.

[9] K. Srisankardara, V. Sethu, E. Ambikairajah, and H. Li, “Front-end for anti spoofing countermeasures in speaker verification: Scattering spectral decomposition,” IEEE J. Sel. Topics Signal Process., vol. 11, no. 4, pp. 632–643, Jun. 2016.

[10] Z. Wu, J. Yamagishi, T. Kimu, C. Hamilci, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, “AVSspoof: The automatic speaker verification spoofing and countermeasures challenge,” IEEE J. Sel. Topics Signal Process., vol. 11, no. 4, pp. 588–604, Jun. 2017.

[11] T. B. Patel and H. A. Patil, “Significance of source-filter interaction for classification of natural vs. spoofed speech,” IEEE J. Sel. Topics Signal Process., vol. 11, no. 4, pp. 644–659, Jun. 2017.

[12] M. M. Kabir, M. F. Mrdida, J. Shin, I. Jahan, and A. Q. Ohi, “A survey of speaker recognition: Fundamental theories, recognition methods and opportunities,” IEEE Access, vol. 9, pp. 79236–79263, 2021.

[13] R. P. Ramachandran, K. R. Farrell, R. Ramachandran, and R. J. Mammon, “Speaker recognition—General classifier approaches and data fusion methods,” Pattern Recognit., vol. 35, no. 12, pp. 2801–2821, Dec. 2002.

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

[15] X. Huang, A. Acero, H.-W. Hon, and H. F. R. By-Reddy, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[16] T. Kinnunen and P. Aliu, “On separating glottal source and vocal tract information in telephony speaker verification,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., Apr. 2009, pp. 4545–4548.

[17] Z. Wu, L. Zhang, and B. Xia, “Robust auditory-based speech feature extraction using independent subspace method,” in Proc. Adv. Cognit. Neurodynamics (ICCN). Dordrecht, The Netherlands: Springer, 2008, pp. 905–909.

[18] M. A. Abd-el-Malek and S. S. Hanna, “The Hilbert transform of cubic splines,” Berlin, Germany: Springer-Verlag, 2014.

[19] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” Speech Commun., vol. 52, no. 1, pp. 12–40, 2010.

[20] Y. Zouhir and K. Ouni, “Feature extraction method for improving speech recognition in noisy environments,” J. Comput. Sci., vol. 12, no. 2, pp. 56–61, Feb. 2016.
M. Campi et al.: Machine Learning Mitigants for Speech Based Cyber Risk

[45] S. Boucheron, O. Bousquet, and G. Lugosi, “Theory of classification: A survey of some recent advances,” ESAIM Probab. Statist., vol. 9, pp. 323–375, Nov. 2005.

[46] D. Huang, J. Unnikrishnan, S. Meyn, V. Veeravalli, and A. Surana, “Statistical SVMs for robust detection, supervised learning, and universal classification,” in Proc. IEEE Inf. Theory Workshop Netw. Inf. Theory, Jun. 2009, pp. 62–66.

[47] Institute of Electrical and Electronics Engineers, “IEEE recommended practice for speech quality measurements,” IEEE Trans. Audio Electroacoustics, vol. 17, no. 3, pp. 225–246, 1969.

[48] A. E. Rosenberg, “Recent research in automatic speaker recognition,” in Advances in Speech Signal Processing, New York, NY, USA: Marcel Dekker, 1992, pp. 701–738.

[49] P. Thaine and G. Penn, “Extracting mel-frequency and bark-frequency cepstral coefficients from encrypted signals,” in Proc. Interspeech, Sep. 2019, pp. 3715–3719.

[50] P. Kumar and S. L. Lahudkar, “Automatic speaker recognition using LPC and MFCC,” Int. J. Recent Innov. Trends Comput. Commun., vol. 3, no. 4, pp. 2106–2109, 2015.

[51] M. Chougala and S. Kuntoji, “Novel text independent speaker recognition using LPC based formats,” in Proc. Int. Conf. Electr., Electron., Opt. Techn. (ICEEOT), Mar. 2016, pp. 510–513.

[52] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O’Shaughnessy, “Multitaper MFCC and PLP features for speaker verification using i-vectors,” Speech Commun., vol. 55, no. 2, pp. 237–251, Feb. 2013.

[53] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, “RASTA-PLP speech analysis,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., vol. 1, Dec. 1991, pp. 121–124.

[54] T. Gulzar, A. Singh, and S. Sharma, “Comparative analysis of LPCC, MFCC and BFCC for the recognition of Hindi words using artificial neural networks,” Int. J. Comput. Appl., vol. 101, no. 12, pp. 22–27, Sep. 2014.

[55] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, and K. A. Lee, “ASVspoof 2019: Future horizons in spoofed and fake audio detection,” 2019, arXiv:1904.05441. [Online]. Available: http://arxiv.org/abs/1904.05441

[56] M. Sahidullah, T. Kinnunen, and C. Hanilçi, A Comparison of Features for Synthetic Speech Detection. International Speech Communication Association, Sep. 2015.

[57] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, “Advances in anti-spoofing: From the perspective of ASVspoof challenges,” APSIPA Trans. Signal Inf. Process., vol. 9, p. e2, Jan. 2020.

[58] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 1, May 1996, pp. 373–376.

[59] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” Speech Commun., vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[60] Y. Tabet and M. Boughnazi, “Speech synthesis techniques. A survey,” in Proc. Int. Workshop Syst., Signal Process. Appl. (WOSPA), May 2011, pp. 67–70.

[61] Y. Saito, S. Takamichi, H. Saruwatari, Y. Saito, S. Takamichi, and H. Saruwatari, “Statistical parametric speech synthesis incorporating generative adversarial networks,” IEEE/ACM Trans. Audio, Speech Language Process., vol. 26, no. 1, pp. 84–96, Jan. 2017.

[62] A. D. V. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” in Proc. 9th Speech Synth. Workshop (ISCA), 2016, p. 125.

[63] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” 2017, arXiv:1703.10135. [Online]. Available: http://arxiv.org/abs/1703.10135

[64] P. Parthia, J. Toveark, G. H. Ik, J. Rozhon, and M. Voznak, “Deep learning serves voice cloning: How vulnerable are automatic speaker verification systems to spoofing trials?” IEEE Commun. Mag., vol. 58, no. 2, pp. 100–105, Feb. 2020.

[65] Y. Qian, F. K. Soong, and Z.-J. Yan, “A unified trajectory tilting approach to high quality speech rendering,” IEEE Trans. Audio, Speech, Language Process., vol. 21, no. 2, pp. 280–290, Feb. 2012.

[66] E. L. Campbell, G. Hernández, and J. R. Calvo, “Feature extraction of automatic speaker recognition, analysis and evaluation in real environment,” in Proc. Int. Workshop Artif. Intell. Pattern Recognit. Cham, Switzerland: Springer, 2018, pp. 376–383.

[67] B. J. Shannon and K. K. Paliwal, “A comparative study of filter bank spacing for speech recognition,” in Proc. Microelectronic Eng. Res. Conf., vol. 41, Nov. 2003, pp. 310–312.

[68] G. K. Liu, “Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech,” 2018, arXiv:1806.09010. [Online]. Available: http://arxiv.org/abs/1806.09010

[69] L. V. D. Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, pp. 2579–2605, Nov. 2008.

[70] J. Yamagishi, T. H. Kinnunen, N. Evans, P. D. Leon, and I. Trancoso, “Introduction to the issue on spoofing and countermeasures for automatic speaker verification,” IEEE J. Sel. Topics Signal Process., vol. 11, no. 4, pp. 585–587, Jun. 2017.

[71] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, “ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge,” in Proc. 16th Annu. Conf. Int. Speech Commun. Assoc., Sep. 2015.

[72] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, “The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection,” Tech. Rep., 2017.

[73] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, “t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification,” 2018, arXiv:1804.09618. [Online]. Available: https://arxiv.org/abs/1804.09618

[74] D. Matrouf, J.-F. Bonastre, and C. Fredouille, “Effect of speech transformation on impostor acceptance,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., vol. 1, May 2006, p. 1.

[75] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Trans. Inf. Syst., vol. 99, no. 7, pp. 1877–1884, 2016.

MARTA CAMPI received the B.Sc. degree in mathematical statistics and data processing (SMID) from the Department of Mathematics, School of Mathematical, Physical, and Natural Sciences, University of Genoa, Italy, the M.Sc. degree in financial econometrics from the Department of Economics and Essex Business School, University of Essex, Colchester, U.K., the M.Res. degree in financial computing from the Department of Computer Science, University College London (UCL), London, U.K., and the M.Phil. degree from the Department of Statistical Science, UCL, where she is currently pursuing the Ph.D. degree. During her Ph.D., she was invited to the Department of Statistical Modeling, The Institute of Statistical Mathematics, Tokyo, Japan, as a Research Fellow, to investigate aspects of cyber-security problems.

GARETH W. PETERS is currently working as a Janet and Ian Dunn Endowed Chair Professor in actuarial science and a Chair Professor in statistics for risk and insurance with the Department of Applied Probability and Statistics, University of California at Santa Barbara. He also holds positions as an Honorary Professor of statistics with University College London, since 2018, an Affiliated Professor of statistics with the University of New South Wales, Australia, since 2015, and an Affiliate Member of the Systemic Risk Center, London School of Economics, since 2014, and Oxford Man Institute (OMI), University of Oxford, since 2013. He has also been an Honorary Professor of statistics with The University of Sydney, Australia, since 2016, and Macquarie University, Australia, since 2018. Previously, he was a Chair Professor for statistics in...
M. Campi et al.: Machine Learning Mitigants for Speech Based Cyber Risk

NOURDDINE AZZAOUI received the Maîtrise Diploma degree in applied mathematics and the D.E.A. degree in statistics and applied probability from Fez University, Morocco, in 1999 and 2002, respectively, and the Ph.D. degree in applied mathematics and statistics from the University of Burgundy, Dijon, France, in 2006. From 2006 to 2010, he was an Assistant Professor (ATER) with Burgundy Mathematics Institute and Paul Painlevé Laboratory of Mathematics, Lille, France, and then a Researcher with the Systems Modeling and Dependability Laboratory, University of Technology of Troyes, where he was with the Statistics and Probability Team. Since September 2010, he has been an Associate Professor (Maître de conférences) with Blaise Pascal Laboratory of Mathematics, Clermont Auvergne University.

TOMOKO MATSUI (Senior Member, IEEE) received the Ph.D. degree from the Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, in 1997. From 1988 to 2002, she was with Nippon Telegraph and Telephone, Tokyo, where she was involved in speaker and speech recognition. From 1998 to 2002, she was with the Spoken Language Translation Research Laboratory, Advanced Telecommunications Research Institute International, Kyoto, Japan, as a Senior Researcher, and was involved in speech recognition. In 2001, she was an Invited Researcher with the Department of Acoustic and Speech Research, Bell Laboratories, Murray Hill, NJ, USA, where she was involved in finding effective confidence measures for verifying speech recognition results. She is currently a Professor with The Institute of Statistical Mathematics, Tokyo, where she is involved in statistical modeling for speech and speaker recognition applications. She was a recipient of the Paper Award of the Institute of Electronics, Information, and Communication Engineers of Japan, in 1993.

* * *

risk and insurance with the Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, where he was also the Director of the Scottish Financial Risk Academy (SFRA). Previously, he held tenured positions at the Department of Statistical Sciences, University College London, U.K., and the Department of Mathematics and Statistics, University of New South Wales, Sydney, Australia. He was an Adjunct Scientist in mathematics, informatics, and statistics with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), from 2009 to 2017. He was also a Nachdiploma Lecturer in machine learning for risk and insurance with the Risk Laboratory, ETH Zürich. He has made in excess of 150 international invited presentations and speaker engagements, including numerous key note presentations. He has delivered numerous professional training courses to c-suite executive level industry professionals as well as numerous central banks. He has published in excess of 150 peer-reviewed articles on risk and insurance modeling, two research text books on operational risk and insurance as well as being an editor and a contributor for three edited text books on spatial statistics and Monte Carlo methods. He is an Elected Member of the Young Academy of Scotland in the Royal Society of Edinburgh (YAS-RSE) and an Elected Fellow of the Institute of Operational Risk (FIOR).