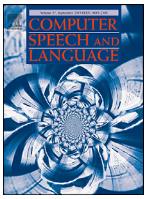
Computer Speech & Language 65 (2021) 101133

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl



Detection of speech playback attacks using robust harmonic trajectories

Wei Shang*, Maryhelen Stevenson*

Dept. of Electrical and Computer Engineering, University of New Brunswick, Fredericton, Canada

ARTICLE INFO

Article History:
Received 13 August 2019
Revised 1 June 2020
Accepted 10 July 2020
Available online 16 July 2020

Keywords:
audio fingerprinting
speech playback/replay attacks
spoofing countermeasure
automatic speaker verification

ABSTRACT

In this paper, a new feature set is proposed for use in a playback attack detector (PAD) aimed at safeguarding a passphrase and speaker-verified protected system that can be remotely accessed from an arbitrary location using an arbitrary telecommunication channel. The new feature set, termed *VoicedTracks*, is a time-frequency map of the most robust harmonic trajectories in an utterance and serves as an *audio fingerprint* that can uniquely identify an utterance despite a moderate amount of noise and channel distortion. Experimental results are obtained using a specially designed in-house database; the impact of various noise types and SNR levels is further investigated using a publicly available database. An analysis of playback scores across several combinations of telecommunication channel types, playback devices and additive noise demonstrates robustness of the feature set to channel distortion and additive noise, thus making it suitable for use in a copy-detection based PAD (cd-PAD) designed for applications such as telephone banking. Relative to other cd-PADs the proposed approach was better able to defend against playback attacks when telephone channels were involved. An analysis of its performance across the replay configurations used in the ASV-spoof 2017 V2 evaluation set suggests that the proposed cd-PAD is highly effective in detecting those playback attacks that are most likely to spoof the speaker verification system.

1. Introduction

In the last decade, the use of biometrics for human authentication has grown to become an important component in many security-related applications worldwide (Caldwell, 2014; Evans et al., 2014b; Xiao, 2007). The advances in communication and computer technologies have allowed rapid adoption and increased effectiveness of both, conventional and novel biometric modalities. However, the success of a biometric modality is not only associated with its effectiveness, but also the availability and affordability of the technology required to measure the human feature(s) it uses (Allano et al., 2006; Jain et al., 2013). Voice, face and fingerprint recognition are the most common examples of the widespread use of biometrics in the market nowadays. But among these modalities, voice biometrics has the advantage that it can be easily captured, not only in person with readily available personal devices, but also remotely with most existing telephone equipment. Hence, voice recognition, in particular speaker verification (Campbell, 1997; Jain et al., 2013), continues to gain the attention of both research groups and industry (Caldwell, 2014; Miller and Fauve, 2012). In 2014, for example, various banking telephone systems around the world considered using spoken passwords for client transactions (Caldwell, 2014).

As with any biometric, automatic speaker verification (ASV) systems are vulnerable to spoofing attacks in which an impostor tries to claim the identity of someone that s/he is not (Evans et al., 2014b; Wu et al., 2014). Depending on the source of the

*Corresponding authors.
E-mail addresses: wei.shang@unb.ca (W. Shang), stevenso@unb.ca (M. Stevenson).

https://doi.org/10.1016/j.csl.2020.101133

utterance used in the attack, spoofing attacks may be categorized as *mimicry attacks*, *synthetic attacks*, and *playback attacks*. In mimicry attacks, the utterance is spoken by a professional mimic or a person with a voice similar to the true client (Hautamäki et al., 2013; Lau et al., 2004). In more sophisticated cases, voice transformation is performed on the impostor's utterance so that it sounds more like the true client (Alegre et al., 2013; Evans et al., 2014a; Kinnunen et al., 2012; Pal and Saha, 2015; Patrick et al., 2005; Wu and Li, 2013). For synthetic attacks, a speaker model is first trained using speech from the true client; the resulting speaker model is then employed to synthesize the utterance to be used in the attack (De Leon et al., 2012; McClanahan et al., 2014; Sanchez et al., 2015; Satoh et al., 2001; Wu et al., 2013). Finally, for playback attacks, a recording of the true client's utterance is used (Alegre et al., 2014b; Galka et al., 2015; Shang and Stevenson, 2008a; Wu et al., 2012; Gonzalez-Rodriguez et al., 2018). Of the three types of spoofing attacks, playback attacks pose the most serious threat, providing an effective means of spoofing the security system while requiring very little technical knowledge or skill to execute (Alegre et al., 2014a; Evans et al., 2014b; Lindberg and Blomberg, 1999; Kinnunen et al., 2017). As a consequence, there is an increasingly pressing need for robust countermeasures to playback attacks.

Traditionally, *prevention techniques* such as the use of a text-prompted ASV system have been suggested as a means of deterring playback attacks (De Leon et al., 2012; Faundez-Zanuy et al., 2006; Hong et al., 2014). While a text-prompted system reduces the risk of a playback attack, it requires additional user interaction thereby compromising the user experience. Furthermore, when compared to a passphrase-protected system for which a passphrase known only to the true client is uttered to gain access, it loses the protection that comes with requiring knowledge of the passphrase, making the text-prompted system more vulnerable to mimicry and synthetic attacks. The use of additional biometric modalities has also been suggested (Bredin et al., 2006; Faundez-Zanuy et al., 2006; Lang and Dittmann, 2007; Nematollahi et al., 2014) as a means of preventing playback attacks; in this case, user experience may be negatively impacted and there is a clear tradeoff between system security and system complexity.

To reduce the susceptibility of passphrase-protected speaker-verified systems to playback attacks, the authors postulate the following list of best practices:

1. *system security log*: all access attempts **must be** logged and their associated audio recordings **must be** stored in the system.
2. *choice of passphrase and reset interval*: the client **must** choose a unique passphrase that is known only to the client, is to be used for the sole purpose of accessing the system, and that satisfies a minimum length requirement; in addition, the passphrase **must be** reset on a regular basis (*e.g.*, annually) or as needed.
3. *means of access*: system access **shall be** carried out in a secure environment with little acoustic interference and, if applicable, via a secure channel from a trust-worthy service provider

While the practicality of the *means of access* best practice is highly dependent on the ASV's exact use case (*e.g.*, it is not entirely practical for telephone banking); it is our position, that the other two can and should be universally applied in the design and implementation of the ASV system. The *system security log* best practice – intended to satisfy modern-day enterprise-level security needs – may be especially important for an application such as telephone banking. Not only will adherence to this best practice provide an information source for the investigation of successfully perpetrated playback attacks, but it will also provide a means by which playback attacks may be detected.

The development of methods to detect playback attacks has attracted the interest of many researchers in recent years (Bredin et al., 2006; Shang and Stevenson, 2008a; 2008b; 2010; Greenhall and Atlas, 2010; Malik, 2012; Villalba and Lleida, 2011; Wang et al., 2011; Wu and Li, 2014; Galka et al., 2015; Yamagishi et al., 2017; Wu et al., 2017; Gonzalez-Rodriguez et al., 2018). One approach focuses on detecting the distortion associated with the recording and playback devices that are used to execute a playback attack (Greenhall and Atlas, 2010; Villalba and Lleida, 2011; Wang et al., 2011; Kinnunen et al., 2017). Another approach is to embed auxiliary information in the utterance (Faundez-Zanuy et al., 2006; Lang and Dittmann, 2007; Nematollahi et al., 2014). For example, Faundez-Zanuy Faundez-Zanuy et al. (2006) proposed that by embedding a time-stamp in the utterance (at the user-end of the telecommunication channel), one will be able to detect playback attacks by validating the embedded information during verification (at the system-end of the telecommunication channel). While these two detection approaches may be effective in certain situations, they may not be suitable for applications such as telephone banking where remote access (from a client-chosen location via a client-chosen telecommunication channel: landline, cellular channel, or VoIP) is desirable. In such applications, all system-end recordings (authentic and playback) have been impacted by the distortion associated with the **unknown** and **different** transmission paths from the user-end to the system-end. Since each transmission path may involve a different telecommunication channel, room configuration, ambient noise, *etc.*, it is all the more challenging to detect the presence of distortion (or lack thereof) from the **unknown** recording/playback devices, *especially* amidst the predominant distortion from the transmission path. Furthermore, when access sites are unsecured and randomly located (as is the case when clients are allowed the convenience of using a personal phone), the time-stamp approach becomes less realistic.

Previous publications suggest that a more suitable approach for applications such as telephone banking is to detect "identical" utterances (Shang and Stevenson, 2008a; Wu et al., 2014; Galka et al., 2015; Gonzalez-Rodriguez et al., 2018). This approach, termed *copy-detection*, takes advantage of the uniqueness of each utterance to detect playback attacks. It assumes that each time a client successfully accesses the system, an *audio fingerprint* (hereon referred to as a feature set) that uniquely identifies the utterance is *stored* by the biometric system. A playback attack is declared whenever the feature set extracted from the incoming recording (*i.e.*, the system-end recording of the incoming utterance), termed the *incoming feature set*, is found to be too similar to one of the *stored feature sets*.

The success of a copy-detection based playback attack detector (cd-PAD) is highly dependent on the choice of feature set. It is worth noting that a feature set appropriate to the task of speaker verification may be poorly suited to the task of playback attack detection. For speaker verification, the features must capture the voice characteristics that are unique to a given speaker so that it is possible to discriminate between utterances of the same phrase by different speakers, while suppressing differences between distinct utterances of the same phrase by the same speaker. In contrast, for playback attack detection, the features must capture characteristics that are unique to a given utterance, so as to allow discrimination between different utterances of the same phrase by the same speaker, while suppressing differences between noisy and channel-distorted versions of the same utterance. In addition, the chosen feature set should be compact so as to decrease both storage requirements and computational burden associated with playback detection.

Typical feature sets used by cd-PAD algorithms are extracted from a time-frequency representation of the recording. The *peakmap* feature set, introduced in Shang and Stevenson (2008a), consists of the time and frequency location of the five highest spectral peaks in each voiced frame of the incoming recording. Thanks to the random nature of the speech production process, peakmaps extracted from two distinct utterances of the same phrase by the same speaker are measurably different. Furthermore, the relatively high magnitudes of the spectral peaks included in the peakmap help to ensure that the peakmap will remain relatively unaffected by noise and distortion. Other spectrogram-based feature sets that have been used include a binary representation of the utterance spectrogram (Wu et al., 2014) and a set of *spectral landmarks* (Galka et al., 2015; Gonzalez-Rodriguez et al., 2018) (*i.e.*, pairs of local maxima from the spectrogram).

In order to better evaluate the PAD performance in the context of remotely accessed applications, the authors collected a specially designed recording database. The database accounts for various factors that may impact the cd-PAD performance, *e.g.*, telecommunication channel type, playback device quality, client and passphrase variation. While satisfactory performance was obtained when using the peakmap feature set, the authors hypothesized that even better performance could be achieved using the harmonic structure of the voiced frames as a basis for selecting a set of peaks to represent the utterance. In particular, the idea is to select representative peaks based on their association with a set of robust harmonic tracks (frequency trajectories of spectral peaks spanning many frames which collectively rise and/or fall in frequency with a consistent ratio in the transition from one frame to the next); selecting peaks in this manner (at the track level instead of at the frame level) should better allow for the inclusion of all utterance-specific spectral peaks of significant magnitude while also reducing the probability of including non-utterance specific peaks due to noise and/or channel distortion.

In this paper, a new feature set termed *VoicedTracks* (VT) is proposed for use in a cd-PAD. The feature extraction process constructs frequency tracks describing the spectral trajectories of the harmonics in the voiced segments of an utterance. The constructed frequency tracks are then used to determine the set of spectral peaks to be included in the VT feature set. As will be shown, this new feature set improves performance over the peakmap feature set (Shang and Stevenson, 2008a; 2008b; 2010), with the equal error rate dropping from 10.14% to 2.26% when a client-and-passphrase independent threshold is used, and the average equal error rate dropping from 5.28% to 1.67% when client-and-passphrase dependent thresholds are used.

The remainder of the paper is organized as follows. In Section 2, assumptions regarding the scenario to which the cd-PAD will be applied are discussed, and the operation of the cd-PAD is reviewed. Implementation details of the cd-PAD algorithm, including the extraction process of the proposed VT feature set, are presented in Section 3, while a description of the UNB (in-house) database is provided in Section 4. In Section 5, the methodology for the performance evaluation is described and the performance of the proposed cd-PAD is compared to two other cd-PADs. In Section 6, further investigations are conducted to analyze the impact of channel distortion and additive noise on the PAD performance. In Section 7, the cd-PAD's performance is evaluated using the publicly available ASVspoof 2017 V2 database (no longer emulating a remote-access scenario). Finally, conclusions are presented in Section 8.

2. The cd-PAD: assumptions, objectives, and operation

To facilitate the description of the cd-PAD algorithm, the various recordings involved in a playback attack (Fig. 1) are first defined: the *intruder's recording* refers to a user-end recording of the original utterance; the *authentic recording* is a system-end recording of a channel-distorted version of the original utterance; and the *playback recording* is a system-end recording of a channel-distorted played-back version of the intruder's recording. As depicted in Fig. 1, all three recordings originate from the same utterance.

Given an incoming (*i.e.*, system-end) recording and a claimed identity, the task of the cd-PAD is to label the incoming recording as *authentic* or *playback*. The cd-PAD makes this determination based on two assumptions: (*i*) all playback attacks are made using an intruder recording that was obtained when the client uttered his/her passphrase during a previous successful system-access attempt; and (*ii*) the system stores the feature sets extracted from all incoming recordings associated with successful access attempts. Provided the preceding two assumptions are valid, as should be the case when adhering to the first two *best practices* postulated in Section 1, any recording used in a playback attack will contain the same utterance as a previous incoming recording whose feature set is stored by the system, thus allowing the playback attack to be detected.

The operation of the cd-PAD algorithm consists of three stages: (A) *feature extraction*: the feature set is extracted from the incoming recording; (B) *similarity assessment*: the incoming feature set is compared to each of the stored feature sets associated with the previous system accesses of the claimed client (assuming a total of N stored feature sets, the result will be a set of N similarity scores); (C) *attack/non-attack classification*: the incoming recording is labelled accordingly as playback or authentic.

In general, the cd-PAD can make two types of classification errors. A *false alarm* occurs when the cd-PAD mislabels an authentic recording as 'playback'; in this case, the cd-PAD incorrectly declares a playback attack and the client is denied access to the

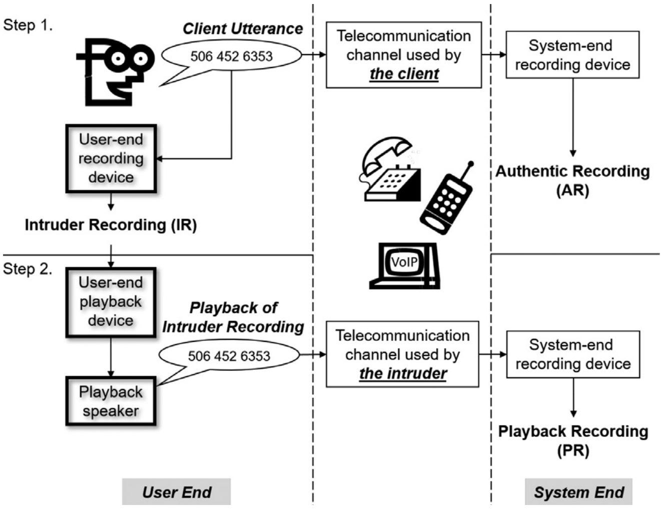

Step 1.
Client Utterance
506 452 6353
Telecommunication channel used by the client
System-end recording device
User-end recording device
Authentic Recording (AR)
Intruder Recording (IR)
Step 2.
User-end playback device
Playback of Intruder Recording
506 452 6353
VoIP
Telecommunication channel used by the intruder
System-end recording device
Playback speaker
Playback Recording (PR)
User End
System End

Fig. 1. Execution of a playback attack

system. A *missed detection* occurs when the cd-PAD mislabels a playback recording as 'authentic'; in this case, the cd-PAD fails to detect a playback attack, resulting in the possibility that an intruder will gain access to the system.

An important advantage of the cd-PAD is the automatic protection it offers against repeated attacks using the same source (*i.e.*, the same intruder recording). In particular, if an intruder gains access to the system due to a missed detection error (*i.e.*, due to insufficient similarity between the intruder's playback and any of the stored recordings), then the feature set extracted from the playback of the intruder's recording will, in practice, be added to the system's collection of stored feature sets - thereby, increasing the odds that any subsequent attack using the same intruder recording will be detected by the cd-PAD. Such automatic protection against repeated attacks using the same source is not offered by systems that do not consider the degree of similarity between the client's current and previous utterances of their passphrase when assessing the legitimacy of a system access attempt.

3. cd-PAD Implementation

Implementation details regarding the three stages of the cd-PAD algorithm, as well as a brief comparison of the VT feature set and the peakmap feature set, are provided in the following subsections. Two authentic recordings, referred to as AR#1 and AR#2, and a playback recording, PR#1, are used to illustrate and compare the extracted feature sets. All three recordings contain utterances of the telephone number "506 452 6353" spoken by the same female speaker (labelled as Client B in the database); recordings AR#1 and PR#1 share the same originating utterance (as depicted in Fig. 1), which is distinct from the utterance in recording AR#2. In addition to the distortion from the transmission channel (cellular channels were used for these three recordings), the utterance in PR#1 also suffers from the distortion of the user-end recording device and playback device (a stereo speaker).

3.1. Extraction of voicedtracks

Extraction of the VT feature set consists of three stages: frame-level peak selection; speech frames detection; and voiced tracks construction. A high-level algorithmic description of the three stages is provided in Table 1. An overview of the extraction process, primarily focused on the principle of the voiced tracks construction process, is provided below; additional details can be found in Shang and Stevenson (2020).

Prior to commencing the voiced tracks construction, the incoming recording (sampled at 8 kHz) is first divided into overlapping Hamming-windowed frames using a frame size of 48 ms and a frame interval of 10 ms. For each frame: the 512-point Fast Fourier Transform (FFT) is computed, spectral peaks are identified, and a peak selection process is undertaken with the goal of retaining all peaks associated with strong harmonics while rejecting weaker peaks including those that may be riding on the sides of the harmonic peaks. Subsequently, all frames are labelled as speech or non-speech and the spectral peaks in the non-speech frames are discarded.

Within each speech segment, (*i.e.*, group of consecutive speech frames), the harmonic frequency trajectories are determined by establishing connections between retained spectral peaks in neighboring frames. For any given frame, the peak matching algorithm first attempts to establish a connection from the highest magnitude peak in the current frame to a peak in the subsequent frame; assuming a connection is established, the algorithm then attempts to establish a connection between the next highest-magnitude peak of the current frame and a previously unmatched peak in the subsequent frame. This peak matching process continues in the current frame until either failing to establish a connection for the current peak or until connections have been established for all retained peaks in the current frame.

Table 1
High-level algorithmic description of the VT extraction process

```
// Stage 1: Preprocessing and frame-level peak selection
import signal from audio file
resample signal at 8 kHz
enframe signal with 48ms frame size and 10ms interval
for each frame:
    apply Hamming window
    compute 512-point FFT
    identify all peaks in the FFT
    sort all peaks in the order of descending magnitude
    for each peak:
      select the peak only if, (and discard otherwise)
        * it is sufficiently wide, and
        * it is of significant magnitude, and
        * it is sufficiently separated in frequency from other selected peaks
    sum the magnitudes of the frame's top five peaks
    calculate the log of the sum of magnitudes (LSM) found in the previous step

// Stage 2: Speech frame detection (energy-based)
generate a histogram of the LSMs from Stage 1
find the threshold that separates the speech frames from the non-speech frames
(the threshold is currently found as the LSM value associated with the first valley
    in the histogram that would yield between 20% and 70% of non-speech frames)
if a threshold is found
   label all frames above threshold as speech frames
else: //no threshold is found
   label all frames as speech if no such threshold is found
eliminate all peaks in non-speech frames

// Stage 3: Voiced tracks construction
for each frame:
    sort peaks in descending order of magnitude
    initialize frame-to-frame frequency ratio bounds, $b_r$, as [0.95, 1.05]
    for each peak: //try to connect to a peak in the next frame
       calculate search interval using FFT bin of current peak and $b_r$
       find the most central peak within the search interval in the next frame
       if a peak is found:
         extend the track by connecting the two peaks
         refine $b_r$ based on the bins of the matched peaks
       else: //no peak is found, and the track ends
         eliminate the track if:
           * it is too short, or
           * its average frequency is outside of desired band
         otherwise:
          record the track
extract locations of all peaks from tracks as the feature set
```

In general, a connection can be established provided there is a retained peak in the next frame that is located within a specific frequency interval (the search interval) and which has not previously been matched to some other peak in the current frame; provided one or more such peaks exists, a connection is established to the peak which is closest to the center of the search interval. When establishing the connection for the first (*i.e.*, highest-magnitude) peak of the current frame, the search interval is initialized to include frequencies between $0.95f_c$ and $1.05f_c$, where f_c, the center of the search interval, is set equal to the frequency of the current-frame peak. When establishing connections for subsequent peaks, an appropriate search interval is determined using the trend of the harmonic trajectory from the current to the next frame, as inferred from previously established connections. For example, if f_m and f_p denote the frequencies associated with the *m*th and *p*th harmonics in the current frame, while f'_m and f'_p denote the same quantities in the next frame, harmonic relationships dictate that $f'_p/f'_m = f_p/f_m$. Hence, assuming the most recently established connection is from the current-frame peak at f_m to the next-frame peak at f'_m, the frequency interval to be searched when attempting to establish a connection for the subsequent current-frame peak, at frequency f_p, will be centered at $(f_p/f_m)f'_m$; the size of the search interval is determined by the uncertainty in the values of f_p, f_m, and f'_m that is introduced by the discrete nature of the FFT.

Once the peak matching algorithm has been applied to all frames in a speech segment, harmonic trajectories (sequences of connected peaks) are identified and examined for robustness. To be considered *robust*, a trajectory must be adequately long[1] so

[1] For the results reported herein, a minimum trajectory length of 5 frames was required.

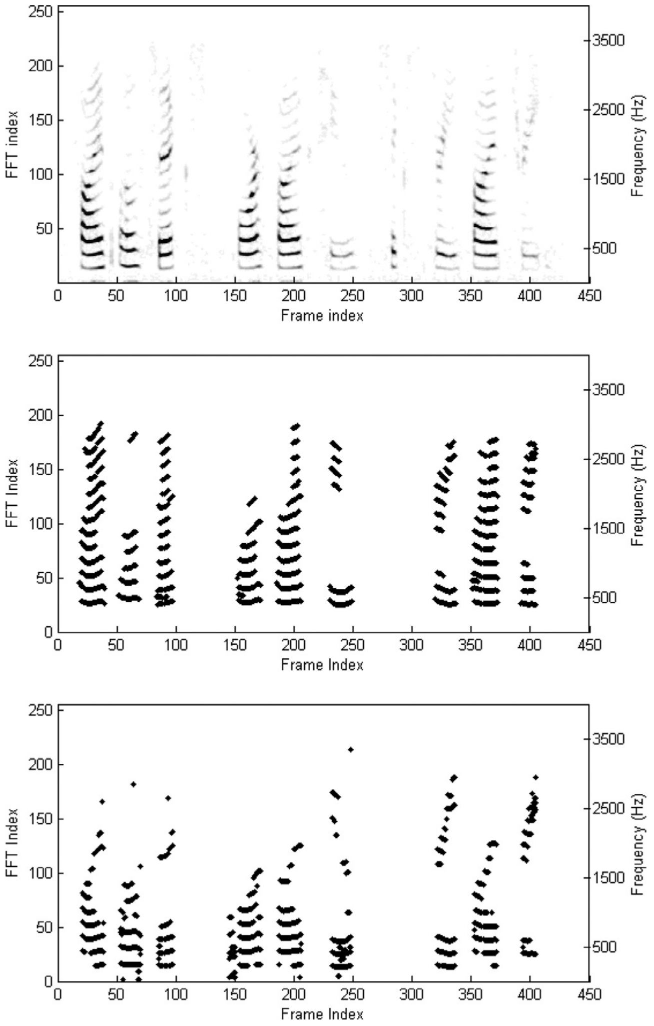


Fig. 2. The STFT (top), the VT feature set (middle), and the peakmap feature set (bottom) of AR#1.

that it is unlikely to have resulted from a string of spurious noise peaks; and it must be contained within a frequency range over which the frequency responses for the telecommunication channels of interest are assumed to be fairly flat.[2]. The criteria for peak selection, frequency track construction, and robustness serve to ensure that the robust frequency tracks will be located within voiced speech frames. The robust frequency tracks are thus, henceforth, referred to as *Voiced Tracks* and the peaks associated with the voiced tracks are called *VT peaks*. The VT feature set is comprised of the relative time and frequency locations of the VT peaks.

To allow for compact storage, easy visualization and efficient comparison of two VT feature sets, the VT feature set may be represented as a binary-valued matrix (henceforth referred to as the *VT matrix*) of size N_f by N_t, where N_f is half the FFT size and N_t is the total number of frames in the utterance. Element (i, j) of the matrix is assigned a value of 1 if the ith FFT value in frame j is a VT peak; otherwise, it is assigned a value of 0. The VT matrix of AR#1 is illustrated in the middle plot of Fig. 2 wherein the elements of the VT matrix having a value of 1 are colored black. For comparison, a gray-scale intensity plot of the STFT magnitude of AR#1 is provided in the top plot of Fig. 2.

3.1.1. *Comparison to the peakmap feature set*

The peakmap feature set is similar to the VT feature set in the sense that it also consists of time-frequency locations of selected spectral peaks from the voiced frames of an utterance. The key difference between the two feature sets lies in the selection of spectral peaks to be included in the feature set; whereas the VT feature set includes all spectral peaks associated with the robust harmonic trajectories, the peakmap feature set includes only the five highest spectral peaks in each of the voiced frames. Fig. 2

[2] For this work, telecommunication channels of interest are typical voice communication channels (*e.g.*, landline, cellular, and VoIP); they are assumed to be adequately flat over the range between 300 and 3000 Hz.

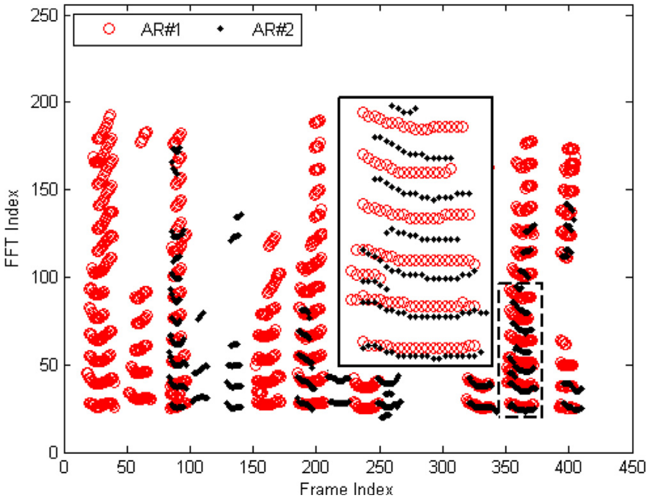


Fig. 3. Comparison of the VT feature sets extracted from authentic recordings AR#1 and AR#2.

compares the peakmap (bottom plot) to the VT matrix (middle plot) of a relatively clean authentic recording AR#1. As seen from the figure, a large percentage of the peaks in the peakmap are also included in the VT matrix (*i.e.*, for most voiced frames, the five highest peaks are also VT peaks). The main exceptions are those peaks in the peakmap feature set that do not belong to a frequency track satisfying the robustness criteria for inclusion in the VT feature set; this includes some high-frequency peaks associated with short-lived harmonic trajectories or noise as well as the low-frequency peaks associated with trajectories below 300 Hz.

In general, however, with the addition of more noise and channel distortion, the VT feature set is shown to be more capable of extracting only the utterance-specific information (*i.e.*, the robust harmonic trajectories) that is crucial to the successful operation of the cd-PAD. This is due to the more sophisticated extraction process of VT which (as further detailed in Shang and Stevenson, 2020) includes:

- a peak matching process that utilizes the harmonic structure in the voiced frames of the speech signal, thereby reducing the possibility of capturing peaks that are not on the harmonic trajectories
- the elimination of the limit on the number of peaks that can be selected for each frame, thereby increasing the pool of spectral peaks for voiced tracks construction
- a more robust energy based speech/non-speech frame detector
- a peak selection process at the track-level, as opposed to the all-or-none frame-level approach used in peakmap

3.2. *Similarity Assessment*

Assuming the claimed client has previously accessed the system a total of N times using his current passphrase, the system will have a total of N stored feature sets representing the associated passphrase utterances. The goal of the similarity assessment stage is to assess the similarity of the incoming feature set with each of the stored feature sets. This is done using the similarity measure described below. The result is a set of N similarity scores with values ranging between 0 (indicating no similarity) and 1 (indicating that the two feature sets are identical).

The first step in determining the similarity score, s_{AB}, of two VT matrices, $A \in \{0,1\}^{N_f \times N_t^A}$ and $B \in \{0,1\}^{N_f \times N_t^B}$, is to find the best alignment of their columns (*i.e.*, the best time alignment). For this purpose, the cross correlation of A and B is evaluated, solely as a function of the column (*i.e.*, frame) displacement variable, ℓ; the row (*i.e.*, frequency) displacement is restricted to a value of zero. The resulting cross-correlation function may be expressed as:

$$r_{AB}(\ell) = \sum_{j=1}^{N_t^A} \mathbf{a}(j) \cdot \mathbf{b}(j-\ell), \quad -N_t^B+1 \le \ell \le N_t^A-1 \tag{1}$$

where: $\mathbf{a}(j)$ denotes the jth column of A, $j = 1, \ldots, N_t^A$; $\mathbf{b}(k)$ denotes the kth column of B for $k = 1, \ldots, N_t^B$ and is defined as an all zero vector of size $N_f \times 1$ for all other k; and $\mathbf{u} \cdot \mathbf{v}$ denotes the dot product of $\mathbf{u}$ and $\mathbf{v}$.

The similarity score, s_{AB}, is then found as the maximum value of the cross-correlation function, normalized, as shown in (2), by the product of the Frobenius norms of A and B. The normalization ensures that similarity scores will never exceed a value of 1 and that a similarity score of 1 will only be obtained in the case that one feature set is identically equal to a frame-shifted version of the other.

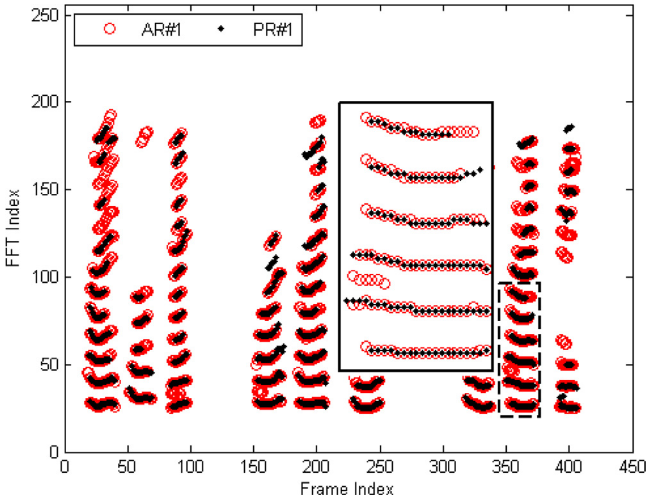


Fig. 4. Comparison of the VT feature sets extracted from authentic recording AR#1 and playback recording PR#1.

$$s_{AB} = \frac{1}{(\| A \|_F)(\| B \|_F)} \max_{\ell} r_{AB}(\ell) \tag{2}$$

where:

$$\| A \|_F = \left(\sum_{j=1}^{N_t^A} \mathbf{a}(j) \cdot \mathbf{a}(j) \right)^{1/2} \text{ and } \| B \|_F = \left(\sum_{k=1}^{N_t^B} \mathbf{b}(k) \cdot \mathbf{b}(k) \right)^{1/2}$$

It is useful to note that the binary-valued nature of A and B greatly simplifies the computation of s_{AB}. In particular, for a frame displacement of ℓ, the value of the cross-correlation is easily determined as the number of 1’s in the element-wise logical AND of the two matrices when aligned so that B is shifted by ℓ columns relative to A. Furthermore, $||A||_F$ is simply equal to the square root of the number of 1’s in A, and similarly for $||B||_F$.

Fig. 3 depicts the VT matrices of AR#1 and AR#2 superimposed on the same plot with the frames of AR#2 displaced relative to the frames of AR#1 so as to illustrate the alignment that produced the maximum value of their cross-correlation function. A similar comparison of the VT matrices of authentic recording AR#1 and playback recording PR#1 (both originating from the same utterance) is shown in Fig. 4. In both figures, a zoom-in of the portion of the VT matrices enclosed by the dashed rectangle (corresponding to the last “5” in the passphrase “506 452 6353”) is provided. As would be expected, there is a better match between the two VT matrices in Fig. 4 than between the two in Fig. 3. The comparison of the VT matrices for AR#1 and AR#2 yields a relatively low similarity score of 0.0626 whereas, the comparison of the VT matrices for AR#1 and PR#1 yields a much higher similarity score of 0.6825.

3.3. Attack/non-attack Classification

The cd-PAD’s labelling decision is based on the *maximum similarity score*, which is defined as the maximum of the N similarity scores that resulted from assessing the similarity of the feature set extracted from the incoming recording with each of the N stored feature sets. If the maximum similarity score is greater than a pre-defined threshold, the incoming recording will be labelled as ‘playback’; otherwise, it will be labelled as ‘authentic’.

4. UNB Database Description

For the purpose of evaluating the performance of the cd-PAD, in the context of a remotely accessed application, we have used an in-house database that was specially designed for such purpose; it was collected by the authors in 2007 at the University of New Brunswick (UNB) and is herein termed the **UNB Database**. Different from other databases that have been used for evaluating the performance of cd-PAD algorithms (Galka et al., 2015; Wu et al., 2014), the utterances captured in all of the authentic and playback recordings have been uniquely distorted via their individual transmission through real publicly-used telecommunication channels of various types – thus mimicking the scenario where the client and the intruder both access the system remotely via their own chosen means of communication.

The database was collected from four participants (hereon referred to as clients A, B, C, and D); clients A and C are male whereas B and D are female. Over a span of four months, each client partook in 90 recording sessions. For each recording session, the client connected to the system, from a location of their choosing, using one of three telecommunication channel types

Table 2
CaP set associated with each client and passphrase pair

Phrase \ Client	A	B	C	D
"506 452 6353"	CaP#1	CaP#2	CaP#3	CaP#4
"University of New Brunswick"	CaP#5	CaP#6	CaP#7	CaP#8
"E3B 5A3"	CaP#9	CaP#10	CaP#11	CaP#12

Table 3
Breakdown of recordings in each CaP set

Recording Type	# of rec.	Set: definition and use	# of rec.
Intruder Recordings	30	**IR**: user-end recordings of client utterances (see Figure 1); used for making playback recordings	30
Authentic Recordings	90	**AR-stored**: system-end recordings of client utterances for which IRs exist; used as the set of stored recordings (the source for the stored feature sets)	30
		AR-eval: system-end recordings of client utterances for which IRs do not exist; used as trial incoming recordings during performance evaluation	60
Playback Recordings	270	**PR-eval**: system-end recordings of IRs played back via acceptable quality devices at the user-end; used as trial incoming recordings during performance evaluation	180
		PR-dvr: system-end recordings of IRs played back via the low-quality built-in DVR speaker at the user end; used in the channel distortion investigation	90
Total		390	

(landline, cellular, or VoIP); participants were instructed to use each channel type for 30 (*i.e.*, one third) of their sessions. Once connected to the system, the participant uttered three different passphrases (a telephone number: *506 452 6353*; a phrase: *University of New Brunswick*; and a postal code: *E3B 5A3*) while holding the channel-input microphone approximately one inch from the side of the mouth. The utterances were recorded at the system end of the telecommunication channel using a computer equipped with a Dialogic D/21H telephony board (8 bits per sample at a sampling rate of 8 kHz). This resulted in a total of 90 authentic recordings for each of the 12 client and passphrase (CaP) combinations - with the 90 authentic recordings being equally divided among the three telecommunication channel types.

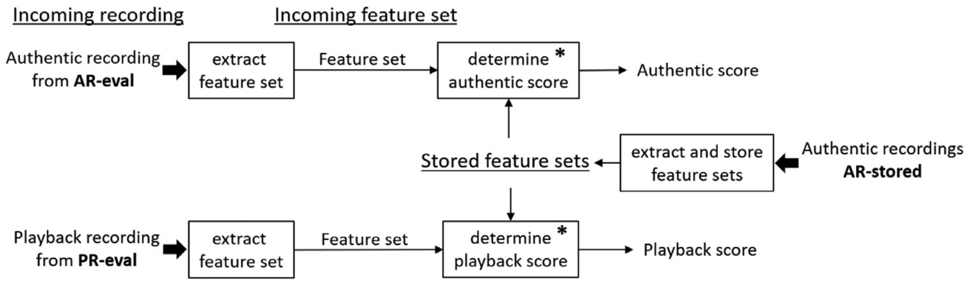
For 30 of the 90 recording sessions (10 sessions for each telecommunication channel type), the client's utterance was simultaneously recorded at the user-end of the channel, using an OLYMPUS WS-100 digital voice recorder (DVR), set at a sampling rate of 44.1 kHz. During these 30 sessions, the participant held two microphones (*i.e.*, they held the DVR microphone in addition to the channel-input microphone) approximately one inch from and on opposite sides of the mouth. These 30 user-end recordings are called intruder recordings and are used for the purpose of making playback recordings.

As illustrated in Fig. 1, playback recordings are made at the system end in the same way as authentic recordings — the only difference occuring at the user end — where, in lieu of the client uttering a passphrase, an intruder recording is played back through a playback device (*e.g.*, a speaker connected to the DVR); the resulting acoustical signal is then captured, as ususal, by the channel-input microphone (positioned approximately one inch in front of the playback device) and transmitted to the system end of the telecommunication channel, where the playback recording is made. Each playback recording is made using one of three telecommunication channel-types (landline, cellular, and VoIP) and one of three playback devices: a stereo speaker (high quality), an external computer speaker (medium quality), and the built-in DVR speaker (very low quality). For each of the 30 intruder recordings, a total of nine playback recordings were made (one using each of the nine possible telecommunication channel/playback-device combinations) resulting in a total of 270 playback recordings for each CaP combination.

Operation of a cd-PAD requires a set of stored recordings (*i.e.*, recordings of the client's passphrase utterance collected during past successful system access attempts); the subset of 30 authentic recordings, for which intruder recordings were simultaneously made, is intended for this purpose – and is henceforth referred to as **AR-stored**. The remaining 60 authentic recordings (*i.e.*, those for which associated intruder recordings were not made) are each used, during performance evaluation, as an *incoming recording* representing a new access attempt by the true client; we refer to this subset of the authentic recordings as **AR-eval**.

Note that each playback recording originates from the same utterance as one of the 30 stored authentic recordings. Despite the same originating utterance, the two recordings contain differently distorted versions of the utterance; whereas the utterance in the stored authentic recording is mainly distorted by the client-chosen telecommunication channel, the utterance captured in the playback recording is distorted by the intruder's: recording device, playback device, and telecommunication channel choice.

In total, the database contains 4680 recordings equally divided among 12 CaP sets. The client and passphrase associated with each CaP set is provided in Table 2; the breakdown and grouping of the recordings are summarized in Table 3. As shown in Table 3, the set of playback recordings is divided into two subsets: **PR-eval** and **PR-dvr**. Similar to the authentic recordings in **AR-eval**, the playback recordings in **PR-eval** will be used as trial incoming recordings during performance evaluation. The subset **PR-dvr** contains the 90 playback recordings that were made using the low-quality built-in DVR speaker; these recordings have a hollow tinny sound and are thus considered to be severely distorted – it is highly unlikely they would be accepted by an ASV


Incoming recording
Incoming feature set
Authentic recording from AR-eval
extract feature set
Feature set
determine authentic score *
Authentic score
Stored feature sets
extract and store feature sets
Authentic recordings AR-stored
Playback recording from PR-eval
extract feature set
Feature set
determine playback score *
Playback score

** The process involves comparing the incoming feature set to each stored feature set and then finding the maximum similarity score.*

Fig. 5. The procedure used to obtain an authentic score for each recording in **AR-eval** (top path) and for obtaining a playback score for each recording in **PR-eval** (bottom path).

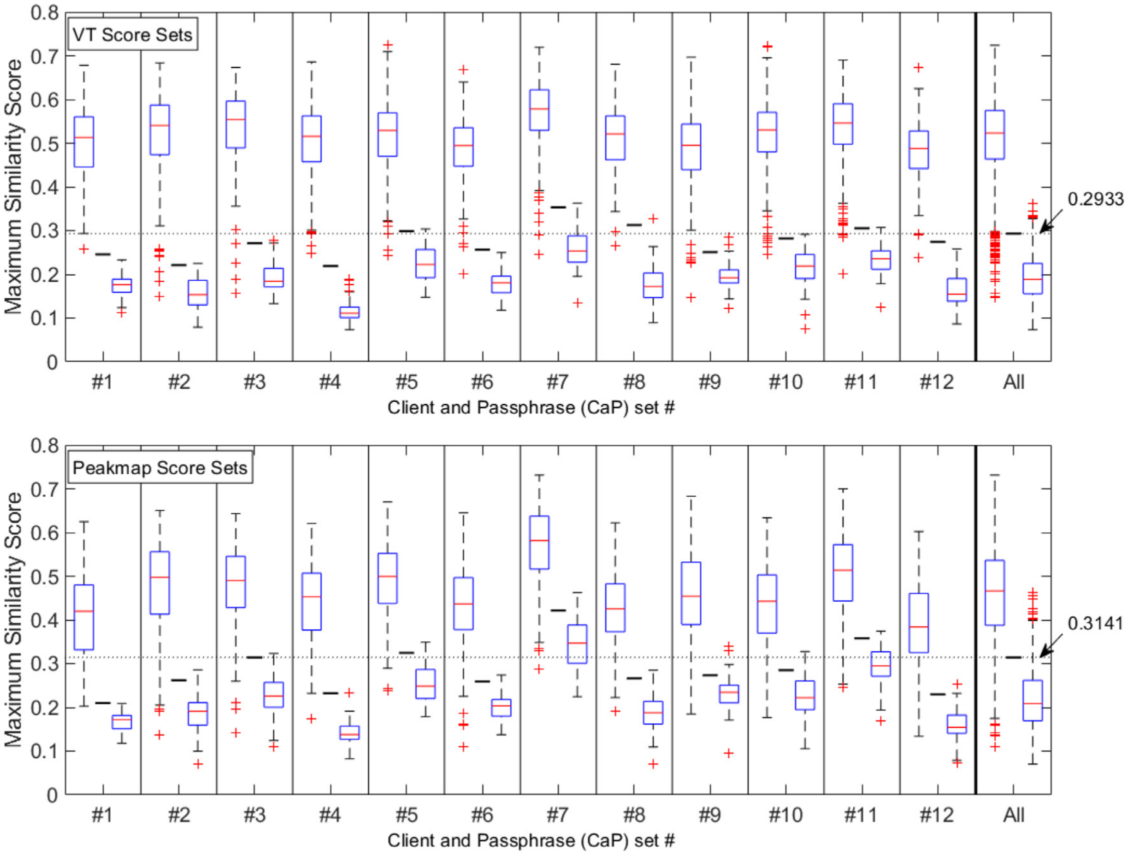


Fig. 6. Box plots of playback and authentic scores obtained for each of 12 CaP sets as well as the pooled scores obtained using the VT (upper) and peakmap (lower) feature sets. Within each CaP section, the box plot of playback scores is offset to the left and the authentic scores to the right. The short solid line, horizontally centered within a CaP section, depicts the CaP-dependent EER threshold, while the dotted line spanning the 12 CaP sections depicts the CaP-independent EER threshold.

system. For this reason, they are not considered to pose a realistic threat and are thus not used for perforsmance evaluation; they will, however, be used in the channel distortion investigation of Section 6.

5. Performance Evaluation and Comparison

This section includes: a description of the methodology used for performance evaluation; a detailed comparison of score sets - illustrating several advantages of the VT feature set relative to the peakmap feature set; DET curves - illustrating and comparing the overall performance of the VT feature set, the peakmap feature set, and the spectral landmarks approach (Galka et al., 2015); and a comparison of the VT and peakmap feature sets with respect to their ability to detect the severely distorted playback attacks that were otherwise excluded from the performance evaluation.

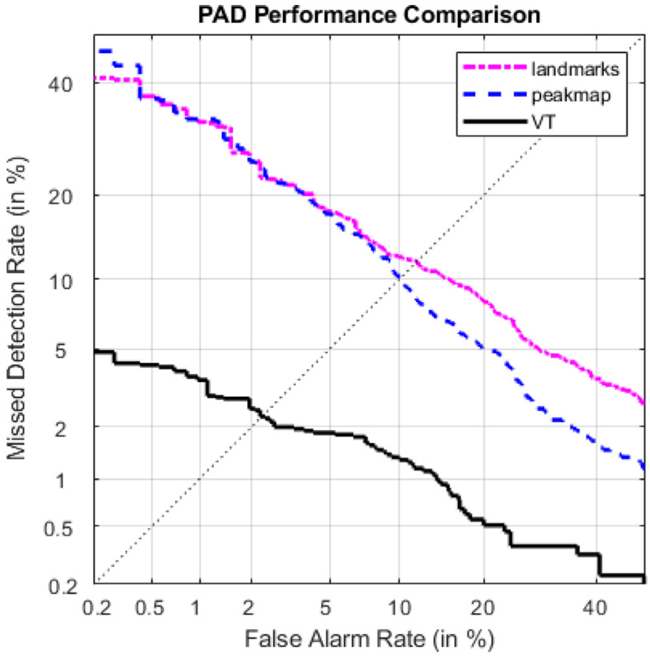


Fig. 7. Detection Error Tradeoff curves for voicedtracks, peakmap, and spectral landmarks. The corresponding EERs are 2.26%, 10.14% and 11.67% respectively.

For each of the 12 CaP sets in the database, the performance of the PAD was evaluated as follows. One by one, each of the 60 authentic recordings (**AR-eval**) and each of the 180 playback recordings (**PR-eval**) was chosen to play the role of *the incoming recording*. As depicted in Fig. 5, the feature set of the incoming recording was compared to each of the 30 stored feature sets (extracted from the recordings in **AR-stored**), resulting in a set of 30 similarity scores. The maximum of the 30 scores was determined and recorded as an authentic score (playback score) when the incoming recording was an authentic recording (playback recording). The process resulted in a set of 60 authentic scores and 180 playback scores for each CaP set.

Fig. 6 compares the box plots (Mcgill et al., 1978) of the playback scores and authentic scores obtained for each of the 12 CaP sets when using the VT feature set (upper plot) to those obtained when using the peakmap feature set (lower plot). For each CaP set, the box plot of the playback score set is offset to the left, while that of the authentic score set is offset to the right. As per MATLAB's boxplot function (MathWorks, 2017), the tops and bottoms of a box represent the 75th and 25th percentiles of the associated score set, while the horizontal line within the box depicts the median. Whiskers extend above and below the box, to the furthest scores within a distance of 1.5 × *IQR*, where *IQR* denotes the *interquartile range*, as depicted by the height of the box. Scores that are considered as outliers, *i.e.*, those that are more than (1.5 × *IQR*) away from the top and bottom of the box, are displayed with a red "+" sign. From Fig. 6, we observe that for each CaP set, there is a significant ***gap*** from the top of the authentic scores' box to the bottom of the playback scores' box, regardless of which feature set is used, thus demonstrating the utility of either feature set for a cd-PAD.

For each CaP set, the *posterior* CaP-dependent equal error rate (EER) threshold was determined as the threshold that yields the minimum difference between the false alarm rate (FAR) and missed detection rate (MDR) and the EER was calculated as the average of the FAR and MDR when using the CaP-dependent EER threshold. In a similar fashion, the *posterior* CaP-independent EER threshold was determined from the ***combined score set*** (a combination of all 12 CaP score sets). Using the CaP-independent EER threshold, the EER for the combined score set was determined as well as the average error rate (AER) for each of the 12 CaP sets, with each AER calculated as the average of the associated FAR and MDR. Note that the average of these 12 AERs is equal to the EER for the combined score set.

When using CaP-dependent EER thresholds (depicted in Fig. 6 by the short horizontal solid lines between the playback scores and authentic scores of each CaP set), the average of the 12 EERs is found to be 1.67% for the VT feature set as compared to 5.28% for the peakmap feature set. In addition to yielding a lower average EER, the VT feature set also offers a more consistent performance across the 12 CaP sets; the standard deviation of the EERs is 1.38% for VT *vs.* 3.09% for peakmap.

The better performance from the VT feature set can be explained by the wider gaps between the boxes of the playback scores and the authentic scores. In particular, the average gap size increases by 66% (from 0.154 when using peakmap to 0.256 when using VT). The increased gap size results from: an increase in the median of each playback score set (indicating an increased robustness against noise and channel distortion); a decrease in the median of each authentic score set (indicating a better capture of the uniqueness of each utterance); and a decrease in the *IQR* of each playback score set (indicating more consistent scores across different channels and playback devices).

Another advantage of the VT feature set can be seen in Fig. 6 as the better alignment and larger overlap of the gaps for the 12 CaP sets. This suggests the potential application of a CaP-independent threshold without triggering significant performance degradation. Indeed, when the CaP-dependent thresholds were replaced by the CaP-independent EER threshold (depicted by the dotted line spanning the width of Fig. 6), the average EER increased by only 35.3% (from 1.67% to 2.26%) for the VT feature set, whereas it increased by 92.05% (from 5.28% to 10.14%) for the peakmap feature set. In addition, the standard deviation of the AERs is 2.75% for VT *vs.* 8.54% for peakmap, indicating that the VT feature set also achieves a more consistent performance when using the CaP-independent threshold.

More generally, the performance improvement, offered by the VT feature set, is illustrated by the detection error tradeoff (DET) curves (Martin et al., 1997) of Fig. 7. The DET curves, which were generated using the combined score set, show the tradeoff between the FAR and MDR as the threshold is varied; the point on the DET curve corresponding to the CaP-independent EER threshold is easily determined as the point where the diagonal dotted line corresponding to MDR=FAR intersects the DET curve; as previously noted, the equal error rate is 2.26% using the VT feature set as compared to 10.14% using peakmap. Also included in Fig. 7 is the DET curve resulting from the spectral landmarks approach (Galka et al., 2015); the implementation adopted by the authors yielded an EER of 11.67% – much higher than that from the VT feature set, but close to that from the peakmap feature set. To the authors' best knowledge, the spectral landmarks approach is the only other cd-PAD that has been suggested for use in a scenario where the client's access to the ASV system is via a telecommunication channel.

The increased robustness to channel distortion, offered by the VT feature set relative to the peakmap feature set, is demonstrated not only by the higher playback scores but also by the cd-PAD's ability to detect severely distorted playback recordings. Comparing the playback scores of the severely distorted playback recordings in the subset **PR-dvr** to the CaP-independent EER threshold for the UNB evaluation set (consisting of the authentic recordings in **AR-eval** and the acceptable quality playback recordings in **PR-eval**), we find that 37.31% of the playback scores obtained from recordings in **PR-dvr** fall below the threshold (and therefore result in missed detections) when using the VT feature set, as compared with 76.11% when using peakmap. Since it is highly likely that the severely distorted playback recordings will be rejected by the ASV system, the detection of such playback attacks is not considered critical to the security of the system; however, with today's heavy focus on security and privacy, knowledge that an attack has occurred still provides tremendous value to clients, service providers and the law enforcement agencies.

6. Investigations

The ability of the feature set to mask differences between recordings that are due to channel distortion or additive noise is key to the successful operation of the cd-PAD. In particular, playback scores must remain high even though the distortion affecting a playback recording may be substantially different than that affecting the associated recording in **AR-stored**. In the following subsections, impacts of channel distortion and additive noise on the playback scores are investigated.

6.1. Impact of channel distortion

For the purpose of this investigation, the playback scores obtained for all playback recordings (in both **PR-eval** and **PR-dvr**) have been divided into 18 groups with each group involving one of the three playback devices and one of the six pairs of telecommunication channels that were used in making the playback recordings and their associated authentic recordings (**AR-stored**). Despite the fact that playback attacks made via the built-in DVR speaker are not considered to pose a realistic threat to the ASV system, the recordings in **PR-dvr** have been included in this investigation so as to better understand the impact of various sources of distortion on the playback scores. Box plots for each of the 18 groups are shown in Fig. 8 for each of the three cd-PAD approaches (VT, Peakmap, and Landmarks) whose performance was evaluated in Section 5. The CaP-independent EER threshold established during the performance evaluation of each approach is indicated by the horizontal dot-dashed line on the corresponding subfigure. Note that in designating the pairs of telecommunication channels, 'C+C' indicates that both the playback recording and its associated stored recording were made via a cellular channel, while 'L+V' indicates that one of the recordings (either the playback or the stored recording) was made via a landline channel and the other made via a VoIP channel, *etc*.

To clearly illustrate the impact of various types of playback devices and telecommunication channels on the playback scores, the 18 groups of playback scores are divided among the three collections of score groups shown in Fig. 8. Collection 3, depicted in the far-right section of the figure, contains the six *DVR-related* score groups, *i.e.*, the six groups of playback scores obtained for the low-quality playback recordings in **PR-dvr** that were made using the built-in DVR speaker as the playback device. Collection 2, depicted in the middle section of Fig. 8, contains the six *VoIP, non-DVR* groups of playback scores; the scores in these groups involve at least one VoIP channel (*i.e.*, either the playback recording, the stored recording, or both contain a version of the utterance that has been transmitted through a VoIP channel) and exclude the built-in DVR speaker as a playback device. Finally, Collection 1 contains the remaining six groups of playback scores, termed the *non-VoIP, non-DVR* score groups. Observations resulting from a comparison of the 18 score groups within the context of the three collections are described and discussed below. Whereas score plots are provided for all three cd-PAD's, our discussion is focused on the playback scores obtained by the VT cd-PAD.

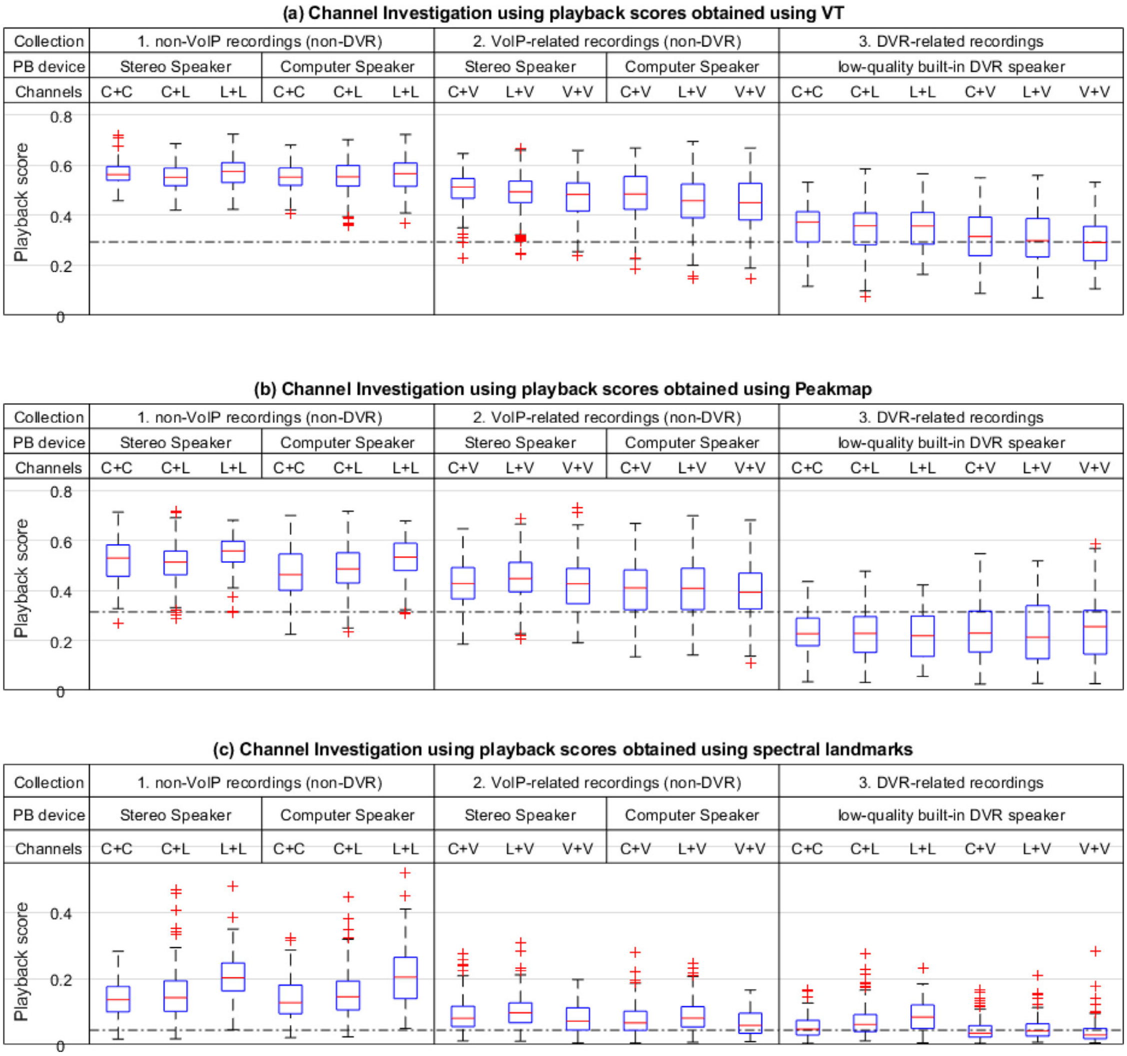


Fig. 8. Box plots of playback scores grouped according to the playback device and telecommunication channel pair for each of three cd-PAD approaches: (a) VT, (b) Peakmap, and (c) Landmarks. The dot-dash line across each subfigure depicts the approach's CaP-independent EER threshold that was obtained from the performance evaluation in Section 5.

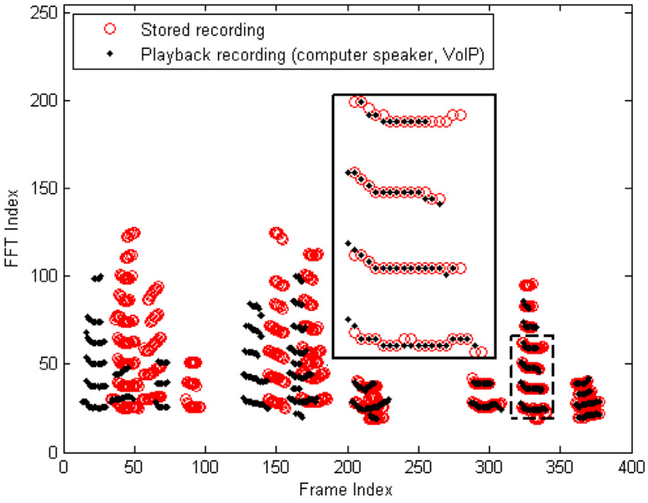


Fig. 9. Comparison of the voicedtracks features sets of a playback recording and its associated stored recording of Client *B* uttering the phone number "506 452 6353".

Observation 1*: Low scores for groups involving the low-quality DVR speaker*

From Fig. 8, it is readily observed that the six groups of playback scores in Collection 3 have medians that are distinctively lower than those in the other collections. The degradation in playback scores is likely due to the significant attenuation of low frequencies by the built-in DVR speaker – which, for the VT cd-PAD approach, results in the dominant peak of a frame occurring at a higher frequency than usual. The higher frequency associated with the dominant peak results in a wider zone (*i.e.*, frequency range) to search when matching harmonic peaks between frames, potentially resulting in the establishment of less accurate harmonic trajectories.

As might be expected, the performance impact introduced by the DVR speaker is greatest for low-pitched speakers. Indeed, of all the playback recordings in **PR-dvr** whose playback scores fell below the VT cd-PAD's CaP-independent threshold established in Section 5, 53.4% were associated with Client A (whose pitch, close to 70 Hz, is quite low); this is significantly higher than the percentage (25%) that would be expected if all four clients were equally affected by the built-in DVR speaker.

Observation 2*: Fairly good scores for the non-DVR groups involving VoIP*

From Fig. 8, it can be observed, for the VT cd-PAD, that the medians of the score groups in Collection 2 are fairly good but slightly lower than those in Collection 1. The lower playback scores for Collection 2 appear to be caused by temporal distortion that is occasionally introduced by the VoIP channel. In particular, the VoIP channel will sometimes change the length of the pauses between the voiced segments, thus making it impossible to simultaneously align all of the corresponding voiced segments in the VT matrices of the playback and stored recordings. This is illustrated in Fig. 9 which compares the VT matrices extracted from a playback recording (computer speaker and VoIP channel) and its associated stored recording (landline), using the alignment that yields the highest similarity score. Despite the fact that these two recordings originate from the same utterance, as evidenced by the well-aligned voiced segments illustrated in the zoomed-in portion from the second half of the utterance, there is visible misalignment between the corresponding voiced segments in the first half of the utterances; the misalignment of these segments results in a low similarity score of 0.2686, which would result in a missed detection if compared to the CaP-independent EER threshold that was established in Section 5.

The distortion introduced by the VoIP channel tends to be worse when the computer speaker is used as a playback device. One possible explanation is that the performance of the speech activity detector, deployed (as a bandwidth savings mechanism) in the front end of the VoIP channel, may be adversely affected by the distortion of the computer speaker; this may result in the VoIP channel wrongfully labelling some of the weaker voiced segments as pauses, which would, in turn, create longer pause sections and more opportunity for temporal distortion.

Observation 3*: Good and consistent scores for the remaining "mixed channel" recordings*

Collection 1 includes the remaining six groups of playback scores. The scores in these groups are associated with playback recordings that were made using acceptable quality playback devices (*i.e.*, the computer speaker or the stereo speaker) in conjunction with a landline or cellular transmission channel; the associated stored recordings are also affiliated with one of these two telecommunication channels. From Fig. 8, it is easily observed that not only are the group medians in Collection 1 the highest, but more interestingly, for the VT cd-PAD, there is relatively little variation among the group medians in this collection. The observed consistency among the score groups in this collection demonstrates the relative insensitivity of the VT feature set to moderate amounts of channel distortion; in particular, differences in the frequency responses of the telecommunication channels (landline *vs.* cellular) and the playback devices (computer speaker *vs.* stereo speaker) have little effect on the playback scores provided the quality of the devices/channels meets a minimum standard.

6.2. Impact of additive noise

The publicly available NOIZEUS database (Hu and Loizou, 2007) was used to examine the cd-PAD's robustness to additive noise. There are 30 noise-free recordings in the database, containing different short phrases, spoken by six speakers. For each of these noise-free recordings, there are a total of 32 noise-added recordings, one for each of the eight noise types[3] at each of the four SNR levels (0, 5, 10 and 15 dB).

To investigate the impact of noise on the cd-PAD's ability to detect playback attacks, the noise-added recordings were compared to their corresponding noise-free recordings and the resulting similarity (*i.e.*, playback) scores were found. Grouped by noise-type and SNR-level, the box plots of the playback scores for each of the 32 (noise-type, SNR-level) groups are provided in Fig. 10, with one subfigure for each of the cd-PAD implementations previously considered. So as to more easily identify box plots associated with the same SNR as well as any trends across the various noise classes, the medians of the score sets, for the same SNR, are connected using one of the four line types indicated in the legend. For the VT cd-PAD, it is interesting to note that, for a given SNR (5 dB or greater), the medians of the playback scores are fairly consistent across most of the noise types (the greatest deviation occurring for the babble noise-type).

As a means of gauging the deterioration in the playback scores (as SNR decreases) and the likelihood that such deterioration will impact the missed detection rate, the CaP-independent EER threshold that was established, using the UNB database, in Section 5 has been depicted as a dashed line across each subfigure. In general, we note that all three cd-PAD's perform reasonably well (the majority of playback scores remaining above the CaP-independent EER threshold) for SNRs of 5 dB and above; for SNR levels lower than 5 dB, the performance of Peakmap is seen to degrade more severely than the other two cd-PAD's. With reference to the VT cd-

[3] The eight noise types are: babble (crowd of people), car, exhibition hall, restaurant, street, airport, train station, train.

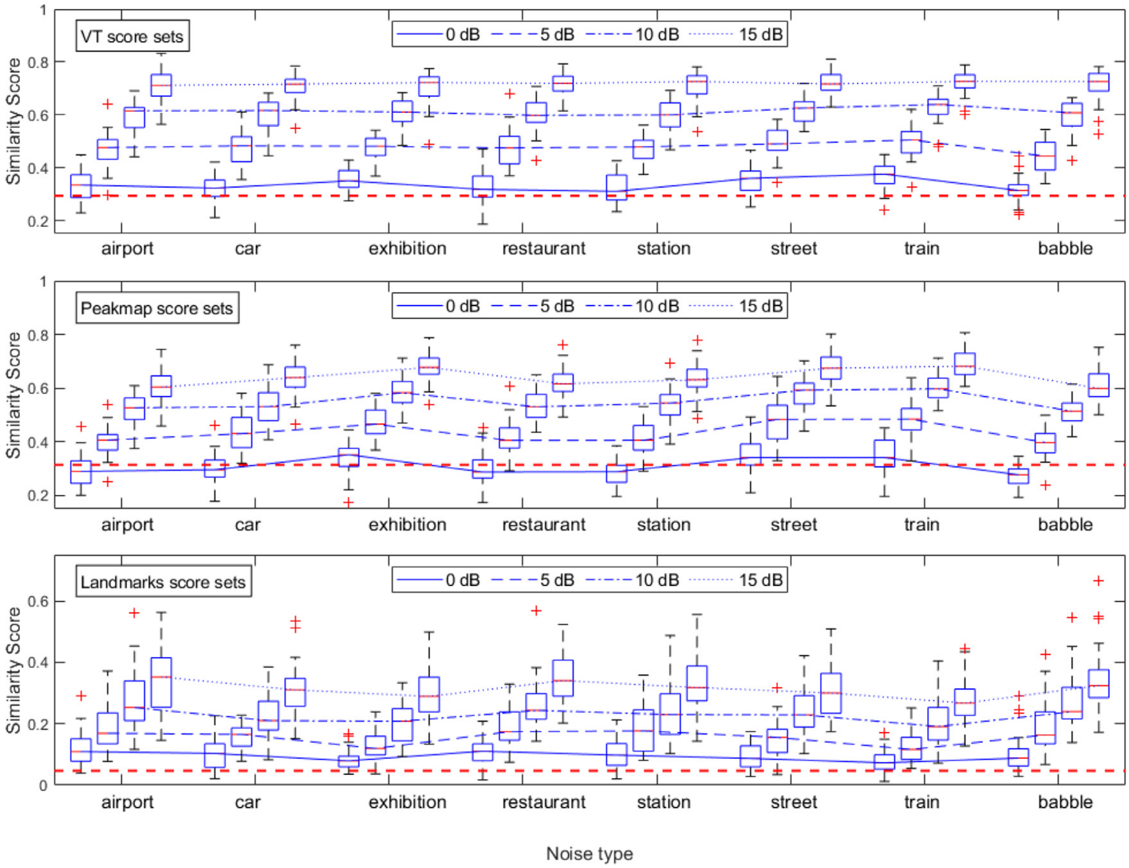


Fig. 10. Box plots of playback score sets resulting from eight noise types at four SNR levels for each of three cd-PAD's: VT (upper), Peakmap (middle), and Landmarks (lower). The dashed line across each subfigure depicts the corresponding cd-PAD's CaP-independent EER threshold for the UNB dataset, as found in Section 5.

PAD, all playback scores are seen to remain above the CaP-independent EER threshold for SNR levels of 5 dB and above, thus demonstrating the ability of the VT feature set to mask differences between recordings due to moderate amounts of additive noise.

7. Experiments using ASVspoof 2017 version 2.0 database

In this section, the performance of the VT cd-PAD is evaluated and compared with other approaches using the ASVspoof 2017 Version 2.0 (Delgado et al., 2018) database. Different from the scenario emulated by the UNB database whereby an intruder recording is played back via a telecommunication channel to attack a *remotely-accessed* passphrase and ASV-protected service (such as telephone banking), the ASVspoof 2017 database emulates a scenario in which an intruder recording is played back to unlock a *nearby* similarly protected smartphone. As a result, the utterances in ASVspoof 2017 are free of distortion from telecommunication channels. Furthermore, the authentic recordings have been re-used as intruder recordings – which differs from the UNB database wherein the utterance captured by an authentic-stored recording is a distorted version of the utterance captured in the corresponding intruder recording. Nonetheless, the quality of the utterances captured in the playback recordings of ASVspoof 2017 is quite varied and allows for additional insight regarding the ability of the cd-PAD to detect playback attacks involving different sources of

Table 4
Statistics of the ASVspoof 2017 Version 2 database .

Set	# of clients	# RC's	# Bona fide	# Spoofed
Training	10	3	1507	1507
Development	8	10	760	950
Evaluation	24	57	1298	12,008

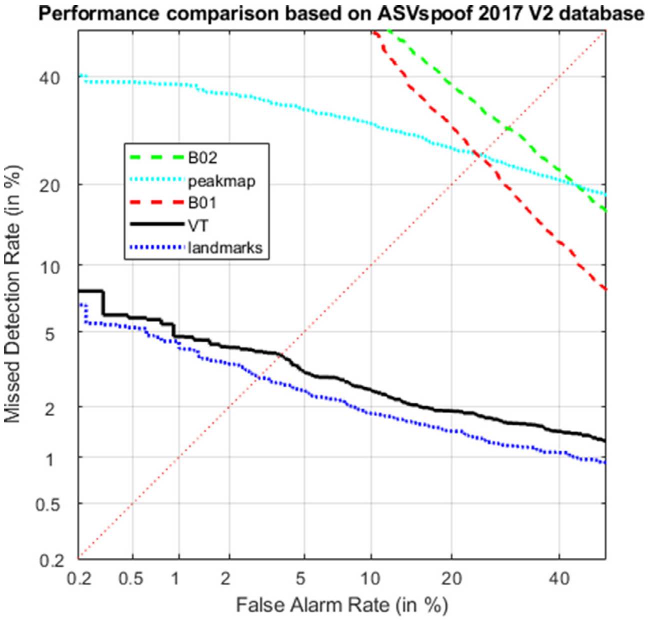


Fig. 11. DET curves showing the performance of five playback attack countermeasures when applied to the ASVspoof 2017 V2 evaluation set.

Table 5
EERs obtained by the various CM approaches in application to the UNB and ASVspoof 2017 Version 2 evaluation sets .

dataset \ CM	VT	Peakmap	Landmarks	B01	B02
ASVspoof 2017 V2	3.83	24.8	2.93	24.4	29.4
UNB	2.26	10.14	11.67	33.6	45.8

distortion as well as additional clients and phrases. Furthermore: the associated ASV threat assessment, provided in Delgado et al. (2018), permits the errors, made by a spoofing countermeasure, to be assessed in the context of the risk they pose to the security of the ASV protected service; and the public availability of the database facilitates comparison with other approaches.

As described in Delgado et al. (2018), Version 2 of the ASVspoof 2017 database has been slightly modified with respect to the original version (Kinnunen et al., 2017) (used in the ASVspoof 2017 competition) so as to remove certain data anomalies. The database consists of *bona fide* (authentic) and *spoofed* (playback) recordings. The bona fide recordings are a subset of the RedDots database (Lee et al., 2015) and the spoofed recordings were generated, using a diverse set of replay configurations (RCs), by playing the bona fide recordings and recapturing their utterances; a characterization of each RC, in terms of its acoustic environment, playback device, and recording device, is provided in Delgado et al. (2018). The database is subdivided into training, development, and evaluation data. The numbers of spoofed and bona fide recordings in each of the training, development, and evaluation sets are provided in Table 4 along with the number of speakers and RC's that are represented in each set.

Since the ASVspoof 2017 corpus was not designed for evaluation of cd-PADs, there is no proper stored recording set. For purposes of evaluating the performance of the cd-PAD algorithms, we have re-used the set of 1298 bona fide evaluation recordings as a stored recording set. To accommodate the fact that each bona fide recording is included in the stored recording set (this will not happen with a proper stored recording set), the score for each bona fide recording is found as the highest similarity score between the bona fide recording being scored and all other bona fide recordings; whereas the score for each playback recording is found in the usual way as the maximum similarity score between the playback recording and all stored recordings. Following this procedure, scores were obtained for the recordings in the ASVspoof 2017 V2 evaluation set using each of the three cd-PAD approaches (VT, peakmap, and spectral landmarks). It is worth noting that the scores produced by the copy-detection based approaches are independent of (*i.e.*, unbiased by) the training and development set. The resulting performances of the cd-PAD approaches are illustrated by the DET curves shown in Fig. 11.

So as to allow for comparison to distortion-detection approaches, the scores for the two baseline CQCC-GMM systems,[4] B01 and B02, of the ASVspoof 2017 competition, were also obtained. The baseline system scores reflect the log likelihood ratio of the bona fide class to the spoofed class, with the class likelihoods being found from appropriately trained Gaussian Mixture Models (GMMs) using feature vectors of constant-Q cepstral coefficients (CQCC). The two baseline systems differ only in the data used for training the GMMs: B01 used the pooled training and development data, while B02 used only the training data.

[4] http://www.asvspoof.org/data2017/baseline_CM.zip.

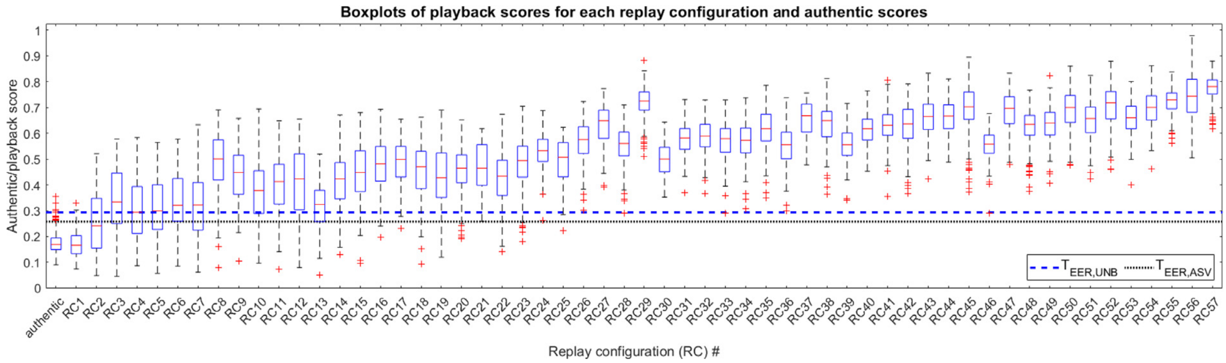


Fig. 12. Box plots of the ASVspoof 2017 V2 scores obtained by the VT cd-PAD. Authentic scores are shown to the far left; playback scores are broken down by replay configuration (RC).

The DET curves for B01 and B02, are illustrated in Fig. 11; it should be noted that their display is consistent with the definitions of false alarm and missed detection used throughout this paper - which differ from those provided in Kinnunen et al. (2017). In particular, the definitions used herein are in accordance with the countermeasure's objective of detecting playback attacks and thus: a missed detection refers to a playback attack that goes undetected by the countermeasure (CM); and a false alarm refers to the case when the CM incorrectly claims that a playback attack occured. The MDR and FAR are calculated using Eqs. (3) and (4).

$$\mathrm{MDR} = \frac{\text{number of playback recordings misclassifed by the CM}}{\text{total number of playback recordings}} \tag{3}$$

$$\mathrm{FAR} = \frac{\text{number of authentic recordings misclassifed by the CM}}{\text{total number of authentic recordings}} \tag{4}$$

As can be seen from the DET curves, the spectral landmarks and VT cd-PAD approaches perform much better than the others. The significant performance advantage offered by VT relative to peakmap highlights the superior ability of the VT extraction process to select robust spectral peaks that capture unique characteristics of the utterance while excluding those that characterize noise. The improved performance of the spectral landmarks approach (relative to its performance on the UNB data set) is likely due to the absence of distortion from real telecommunication channels, which as shown in Fig. 8, tends to impact the scores obtained from the spectral landmarks cd-PAD more so than the scores obtained from the VT cd-PAD. The spectral landmarks approach also seems to benefit from higher-quality stored recordings from which a better set of spectral landmarks can be identified. Finally, the superior performances of the two best performing cd-PAD approaches with respect to the distortion-detection based approaches, illustrates that the cd-PAD has much to offer when safeguarding a system against playback attacks.

Table 5 summarizes the EERs obtained by the various CM approaches in their application to both the UNB and ASVspoof 2017 V2 evaluation sets. For completeness, the Table includes the EERs that were obtained when the baseline systems B01 and B02, trained using the ASVspoof 2017 V2 data, were applied to resampled (16 kHz versions) of the UNB recordings[5]. One possible explanation as to why the distortion-detection based systems (B01 and B02) do not generalize well to the recordings in the UNB database is that, for the UNB database, both playback and authentic recordings contain utterances that have been distorted via their individual transmission through different type telecommunication channels; this is in addition to the distortion due to the playback device which affects only the playback recordings.

So as to gain further insight regarding the impact of attack quality on the VT cd-PAD's ability to detect playback attacks, performance variations across the 57 RCs, identified and numbered in Delgado et al. (2018), were examined. The RC number reflects the rank of the RC in regards to the level of threat its playback recordings pose to the ASV; the threat level of an RC was determined by the EER of an ASV system when zero-effort impostor trials were replaced by playback recordings generated by the RC. The lower threat (*i.e.*, lower numbered) RCs tend to be associated with higher amounts of distortion and hence result in lower-quality playback attacks, which are less likely to be accepted by the ASV as the target speaker.

Fig. 12 depicts the box plots of the playback scores (*i.e.*, the scores obtained for the spoofed recordings) associated with each of the 57 RCs. Also included, to the left of RC1, is a box plot of all authentic (*i.e.*, bona fide) scores; recall bona fide recordings do not involve a RC. In general, good performance is possible when the playback scores of the VT cd-PAD are high, relative to the authentic scores. A missed detection occurs when a playback score falls below the cd-PAD's threshold, whereas a false alarm occurs when an authentic score is higher than the threshold. The score boxplots in Fig. 12 clearly indicate the potential for excellent performance by the VT cd-PAD with respect to the high-threat RCs.

[5] We have also tried retraining B01 and B02, using the same ASVspoof 2017 recordings (but resampled to 8 kHz); application of these retrained B01 and B02 systems to the 8 kHz sampled UNB recordings resulted in EERs of 40.9% and 41.7% respectively.

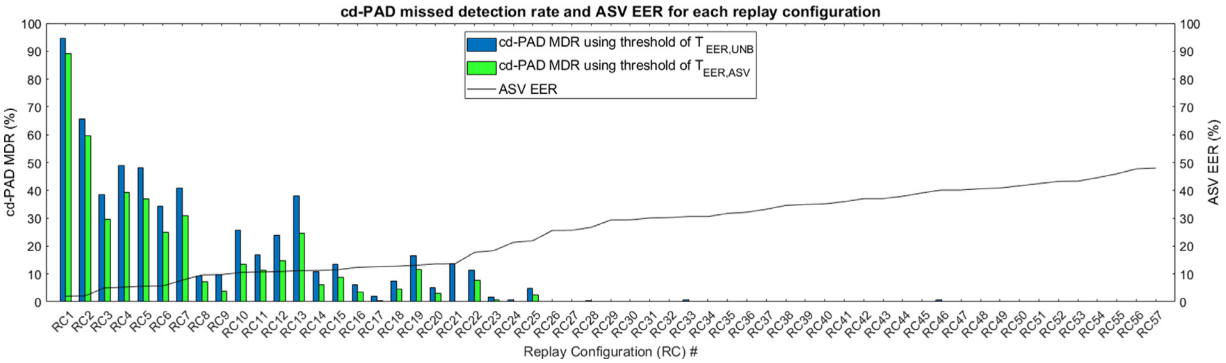


Fig. 13. VT cd-PAD MDRs obtained for each of the 57 RCs when using the two threshold values, $T_{\text{EER,UNB}}$ (blue bars) and $T_{\text{EER,ASV}}$ (green bars), depicted in Fig. 12. Also shown, is the ASV EER (solid line), found and used in Delgado et al. (2018), to assess the threat that playback recordings of each RC pose to the ASV. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To allow for a more quantitative analysis of the impact of RC on the performance of the VT cd-PAD, we consider performances resulting from the two threshold values[6] depicted by the horizontal lines in Fig. 12; these two thresholds, $T_{\text{EER,UNB}}$ (upper line) and $T_{\text{EER,ASV}}$ (lower line), correspond to the VT cd-PAD's EER thresholds for the scores obtained from the UNB and ASVspoof 2017 evaluation sets respectively. (Note that $T_{\text{EER,UNB}}$ is similarly depicted in Fig. 8(a).) In application to the VT cd-PAD scores for the ASVspoof 2017 V2 evaluation set, the threshold $T_{\text{EER,ASV}}$ results in a FAR of 3.93% and a MDR of 3.81% while the threshold $T_{\text{EER,UNB}}$ results in a FAR of 0.77% and a MDR of 5.49%.

Fig. 13 depicts the MDRs that were obtained as a result of applying each of the two threshold values, $T_{\text{EER,UNB}}$ and $T_{\text{EER,ASV}}$, to the playback scores associated with each of the 57 RCs used in the ASVspoof 2017 V2 evaluation set; FAR values are independent of RC. As will always be the case, the lower threshold value, $T_{\text{EER,ASV}}$, results in smaller MDRs than does the higher threshold, $T_{\text{EER,UNB}}$; however, it does so at the cost of a higher FAR (3.93% *vs.* 0.77%). More important is the observation that for both thresholds, the MDR is quite low (and often zero-valued) for the high-threat RCs (say those numbered 23 and higher), whose associated playback recordings have a good chance of successfully spoofing the ASV.

Also shown in Fig. 13, is a plot of the ASV EERs that were obtained and used in Delgado et al. (2018) for ranking the RCs in terms of their threat level. As evident from the Figure, there is a strong negative correlation between the performances of the ASV system and the VT cd-PAD – which not only highlights the crucial role to be played by the cd-PAD in securing a passphrase and ASV protected system against medium-to-high quality playback attacks, but also attests to the complementary roles of the two systems in collectively countering playback attacks of wide-ranging quality. Although the VT cd-PAD may fail to detect the poor quality playback attacks, it is unlikely that these attacks will successfully spoof the ASV; and although the good quality playback attacks may successfully spoof the ASV, they will likely be detected by the VT cd-PAD. In either case, one of the two systems will thwart the attack, thus preventing the perpetrator from gaining access to the protected service.

8. Conclusions

In this paper, the VT feature set has been introduced and proposed for use in a copy-detection based playback attack detector aimed at safeguarding a speaker-verified pass-phrase protected system that can be accessed remotely via a user-chosen telecommunication channel. Different from previous feature sets comprised of time-frequency locations of spectral peaks, the VT feature set includes only those peaks associated with the robust harmonic tracks of the speech signal.

The performance of the feature set was evaluated using the UNB database specifically designed for the intended application. A comparison of score sets revealed several advantages of the VT feature set relative to the previously published peakmap feature set. In particular, the VT feature set was better able to: (1) disregard channel distortion and noise, (2) capture the uniqueness of an utterance; and (3) accommodate the use of a global threshold – thereby eliminating the burdensome task of collecting threshold-setting data for each client, passphrase, playback device, telecommunication channel combination, *etc.* A comparison of DET curves and associated EERs indicated a significant performance advantage of the VT cd-PAD relative to other cd-PAD implementations.

An investigation of playback scores across transmission channels and playback devices served to further demonstrate the insensitivity of the VT feature set to moderate amounts of distortion. A second investigation using NOIZEUS database demonstrated the insensitivity of the VT feature set to moderate amounts of additive noise as well as the consistency of the resulting playback scores across a variety noise types.

[6] These threshold values were chosen for illustration purposes only; in practice, the threshold value should be appropriately chosen for the countermeasure's intended use.

The performance of the VT cd-PAD with respect to the publicly available ASVspoof 2017 V2 database served to demonstrate its versatility - both in terms of its applicability to different types of attack scenarios (in-person as opposed to remote) as well as its ability to handle different types of distortion. The range of VT cd-PAD scores obtained for playback recordings associated with medium-to-high quality playback devices were consistent with those obtained for playback recordings in the UNB database using similar quality playback devices. Finally, the performance of the VT cd-PAD across playback recordings, grouped by level-of-threat posed to the ASV, confirmed the excellent performance of the VT cd-PAD in detecting those attacks to which the ASV is most vulnerable.

Declaration of Competing Interest

We have no conflict of interest to declare.

Acknowledgment

The authors would like to thank both the reviewers and Dr. Julian Cardenas for their many suggestions and valuable contributions to the manuscript. This work was supported, in part, by the Natural Sciences and Engineering Research Council of Canada (NSERC) PGS-D scholarship [PGS D3 - 363217 - 2008].

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.csl.2020.101133.

Appendix A. Supplementary materials

Supplementary Data S1. Feature extraction process for the VoicedTracks feature set. This is open data under the CC BY license http://creativecommons.org/licenses/by/4.0/

References

Alegre, F., Amehraye, A., Evans, N., 2013. Spoofing countermeasures to protect automatic speaker verification from voice conversion. Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 3068–3072. https://doi.org/10.1109/ICASSP.2013.6638222.

Alegre, F., Evans, N., Kinnunen, T., Wu, Z., Yamagishi, J., 2014. Anti-spoofing: voice databases. In: Li, S.Z., Jain, A.K. (Eds.), Encyclopedia of Biometrics. Springer US, pp. 1–7. https://doi.org/10.1007/978-3-642-27733-7_9048-2.

Alegre, F., Janicki, A., Evans, N., 2014. Re-assessing the threat of replay spoofing attacks against automatic speaker verification. Biometrics Special Interest Group (BIOSIG), 2014 International Conference of the. IEEE, pp. 1–6.

Allano, L., Morris, A.C., Sellahewa, H., Garcia-Salicetti, S., Koreman, J., Jassim, S., Ly-Van, B., Wu, D., Dorizzi, B., 2006. Nonintrusive multibiometrics on a mobile device: a comparison of fusion techniques. In: Proc. SPIE, 6202, . https://doi.org/10.1117/12.666088.

Bredin, H., Miguel, A., Witten, I., Chollet, G., 2006. Detecting replay attacks in audiovisual identity verification. In: Proc. ICASSP 2006, 1, p. I.

Caldwell, T., 2014. 2014 a year in biometrics. Biom. Technol. Today 2014 (11), 9–11. https://doi.org/10.1016/S0969-4765(14)70180-8.

Campbell Jr., J.P., 1997. Speaker recognition: a tutorial. Proc. IEEE 85 (9), 1437–1462.

De Leon, P., pucher, M., Yamagishi, J., Hernaez, I., Saratxaga, I., 2012. Evaluation of speaker verification security and detection of hmm-based synthetic speech. IEEE Trans. Audio Speech Lang. Process. PP (99), 1.

Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., Lee, K.A., Yamagishi, J., 2018. ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements. . https://doi.org/10.21437/Odyssey.2018-42.

Evans, N., Alegre, F., Wu, Z., Kinnunen, T., 2014. Anti-spoofing: voice conversion. In: Li, S.Z., Jain, A.K. (Eds.), Encyclopedia of biometrics. second ed. Springer. https://doi.org/10.1007/978-3-642-27733-7_9111-2.

Evans, N., Kinnunen, T., Yamagishi, J., Wu, Z., Alegre, F., De Leon, P., 2014. Speaker recognition anti-spoofing. In: Marcel, S., Li, S., Nixon, M. (Eds.), Handbook of Biometric Anti-spoofing. Springer,. https://doi.org/10.1007/978-1-4471-6524-8_7.

Faundez-Zanuy, M., Hagmüller, M., Kubin, G., 2006. Speaker verification security improvement by means of speech watermarking. Speech Commun. 48 (12), 1608–1619.

Galka, J., Grzywacz, M., Samborski, R., 2015. Playback attack detection for text-dependent speaker verification over telephone channels. Speech Commun. 67 (0), 143–153.

Gonzalez-Rodriguez, J., Escudero, A., de Benito-Gorrn, D., Labrador, B., Franco-Pedroso, J., 2018. An audio fingerprinting approach to replay attack detection on asvspoof 2017 challenge data. In: Proc. Odyssey 2018 The Speaker and Language Recognition Workshop, pp. 304–311. https://doi.org/10.21437/Odyssey.2018-43.

Greenhall, A., Atlas, L., 2010. Cepstral mean based speech source discrimination. In: Proc. ICASSP 2010, pp. 4490–4493.

Hautamäki, R.G., Kinnunen, T., Hautamäki, V., Leino, T., Laukkanen, A.-M., 2013. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In: Proc. INTERSPEECH, pp. 930–934.

Hong, Q., Wang, S., Liu, Z., 2014. A robust speaker-adaptive and text-prompted speaker verification system. In: Sun, Z., Shan, S., Sang, H., Zhou, J., Wang, Y., Yuan, W. (Eds.), Biometric Recognition. Lecture Notes in Computer Science. 8833, Springer International Publishing, pp. 385–393. https://doi.org/10.1007/978-3-319-12484-1_43.

Hu, Y., Loizou, P.C., 2007. Subjective comparison and evaluation of speech enhancement algorithms. Speech Communication 49 (7), 588–601. https://doi.org/10.1016/j.specom.2006.12.006.Speech Enhancement

Jain, A.K., Bolle, R., Pankanti, S., 2013. Biometrics: Personal Identification in Networked Society. Springer Publishing Company, Incorporated.

Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., Lee, K.A., 2017. The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection. INTERSPEECH 2017, Annual Conference of the International Speech Communication Association, August 20–24, 2017, Stockholm, Sweden. Stockholm, SWEDEN.

Kinnunen, T., Wu, Z.-Z., Lee, K.A., Sedlak, F., Chng, E.S., Li, H., 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech. In: Proc. ICASSP 2012.

Lang, A., Dittmann, J., 2007. Digital Watermarking of Biometric Speech References: Impact to the EER System Performance. 6505, . https://doi.org/10.1117/12.703890.

Lau, Y.W., Wagner, M., Tran, D., 2004. Vulnerability of speaker verification to voice mimicking. In: Proc. 2004 Int. Symp. Intelligent Multimedia, Video and Speech Process., pp. 145–148.

Lee, K.-A., Larcher, A., Wang, G., Kenny, P., Brümmer, N., van Leeuwen, D.A., Aronowitz, H., Kockmann, M., Vaquero, C., Ma, B., Li, H., Stafylakis, T., Alam, M.J., Swart, A., Perez, J., 2015. The RedDots data Collection for Speaker Recognition. INTERSPEECH.

Lindberg, J., Blomberg, M., 1999. Vulnerability in speaker verification - a study of technical impostor techniques. In: Proc. European Conf. Speech Commun. and Technology, 3, pp. 1211–1214.

Malik, H., 2012. Securing speaker verification system against replay attack. Audio Engineering Society Conference: 46th International Conference: Audio Forensics.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The det curve in assessment of detection task performance. In: Proc. Eurospeech '97, pp. 1895–1898.

MathWorks, 2017. Box plots.

McClanahan, R., Stewart, B., De Leon, P., 2014. Performance of i-vector speaker verification and the detection of synthetic speech. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 3779–3783. https://doi.org/10.1109/ICASSP.2014.6854308.

Mcgill, R., Tukey, J.W., Larsen, W.A., 1978. Variations of box plots. Am. Stat. 32 (1), 12–16. https://doi.org/10.1080/00031305.1978.10479236.

Miller, D., Fauve, B., 2012. Mobile e-commerce to drive voice-based authentication. Biom. Technol. Today 2012 (2), 5–8. https://doi.org/10.1016/S0969-4765(12)70053-X.

Nematollahi, M., Al-Haddad, S., Doraisamy, S., Ranjbari, M., 2014. Digital speech watermarking for anti-spoofing attack in speaker recognition. Region 10 Symposium, 2014 IEEE, pp. 476–479. https://doi.org/10.1109/TENCONSpring.2014.6863080.

Pal, M., Saha, G., 2015. On robustness of speech based biometric systems against voice conversion attack. Appl. Soft Comput. 30 (0), 214–228. https://doi.org/10.1016/j.asoc.2015.01.036.

Patrick, P., Aversano, G., Blouet, R., Charbit, M., Chollet, G., 2005. Voice forgery using alisp: indexation in a client memory. Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on, 1, pp. 17–20. https://doi.org/10.1109/ICASSP.2005.1415039.

Sanchez, J., Saratxaga, I., Hernaez, I., Navas, E., Erro, D., Raitio, T., 2015. Toward a universal synthetic speech spoofing detection using phase information. Inf. Forensics Secur. IEEE Trans. 10 (4), 810–820. https://doi.org/10.1109/TIFS.2015.2398812.

Satoh, T., Masuko, T., Kobayashi, T., Tokuda, K., 2001. A robust speaker verification system against imposture using an HMM-based speech synthesis system. INTERSPEECH, pp. 759–762.

Shang, W., Stevenson, M., 2008. A playback attack detector for speaker verification systems. In: Proc. 3rd Int. Symp. Commun., Control and Signal Process., 2008. ISCCSP 2008, pp. 1144–1149.

Shang, W., Stevenson, M., 2008. A preliminary study of factors affecting the performance of a playback attack detector. Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on, pp. 000459–000464. https://doi.org/10.1109/CCECE.2008.4564576.

Shang, W., Stevenson, M., 2010. Score normalization in playback attack detection. In: Proc. ICASSP 2010, pp. 1678–1681. https://doi.org/10.1109/ICASSP.2010.5495503.

Shang, W., Stevenson, M., 2020. Feature Extraction Process for the VoicedTracks Feature Set. Technical Report. Department of Electrical and Computer Engineering, University of New Brunswick.

Villalba, J., Lleida, E., 2011. Detecting replay attacks from far-field recordings on speaker verification systems. In: Proc. COST 2011 European Conf. Biometrics and ID Management, pp. 274–285.

Wang, Z.-F., Wei, G., He, Q.-H., 2011. Channel pattern noise based playback attack detection algorithm for speaker recognition. Machine Learning and Cybernetics (ICMLC), 2011 International Conference on, 4, pp. 1708–1713. https://doi.org/10.1109/ICMLC.2011.6016982.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H., 2014. Spoofing and countermeasures for speaker verification: a survey. Speech Commun.. https://doi.org/10.1016/j.specom.2014.10.005.

Wu, Z., Gao, S., Cling, E.S., Li, H., 2014. A study on replay attack and anti-spoofing for text-dependent speaker verification. Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA), pp. 1–5. https://doi.org/10.1109/APSIPA.2014.7041636.

Wu, Z., Kinnunen, T., Chng, E.S., Li, H., Ambikairajah, E., 2012. A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case. Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1–5.

Wu, Z., Li, H., 2013. Voice conversion and spoofing attack on speaker verification systems. Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific, pp. 1–9. https://doi.org/10.1109/APSIPA.2013.6694344.

Wu, Z., Li, H., 2014. Voice conversion *versus* speaker verification: an overview. APSIPA Trans. Signal Inf.Process. 3, e17.

Wu, Z., Xiao, X., Chng, E.S., Li, H., 2013. Synthetic speech detection using temporal modulation feature. Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 7234–7238. https://doi.org/10.1109/ICASSP.2013.6639067.

Wu, Z., Yamagishi, J., Kinnunen, T., Hanili, C., Sahidullah, M., Sizov, A., Evans, N., Todisco, M., Delgado, H., 2017. ASVspoof: the automatic speaker verification spoofing and countermeasures challenge. IEEE J. Sel. Top. Signal Process. 11 (4), 588–604. https://doi.org/10.1109/JSTSP.2017.2671435.

Xiao, Q., 2007. Technology review - biometrics-technology, application, challenge, and computational intelligence solutions. Comput. Intell. Mag. IEEE 2 (2), 5–25. https://doi.org/10.1109/MCI.2007.353415.

Yamagishi, J., Kinnunen, T.H., Evans, N., Leon, P.D., Trancoso, I., 2017. Introduction to the issue on spoofing and countermeasures for automatic speaker verification. IEEE J. Sel. Top. Signal Process. 11 (4), 585–587. https://doi.org/10.1109/JSTSP.2017.2698143.