# A Playback Attack Detector for Speaker Verification Systems

Wei Shang
Department of Electrical and Computer Engineering
University of New Brunswick
Fredericton, Canada, E3B 5A3
Email: wei.shang@unb.ca

Maryhelen Stevenson
Department of Electrical and Computer Engineering
University of New Brunswick
Fredericton, Canada, E3B 5A3
Email: stevenso@unb.ca

***Abstract*—A playback attack detector (PAD), which can be mobilized in guarding speaker verification systems against playback attacks, is described in this paper. To detect playback attacks, the PAD uses a feature set called peakmap, which includes the frame and FFT bin numbers of the five highest spectral peaks from each of the voiced frames in an utterance. During the detection, the peakmap of the incoming recording is first extracted and then compared to those of all the other recordings that are stored at the system end. Each comparison will yield a similarity score that represents the level of similarity between the two recordings. The incoming recording is declared to be a playback recording if its maximum similarity score is above a threshold.**

## I. Introduction

Systems which allow clients the convenience of remote interaction via telephone have become increasingly popular. To safeguard against intruders, clients are typically required to enter a personal identification number (PIN) known only to the client. Such PIN-protected systems can be compromised if a client's PIN is obtained by an intruder. A speaker-verified pass-phrase protected system offers an increased level of security with similar client convenience. In order to spoof such a system, not only does an intruder require knowledge of a client's pass phrase, but the intruder must also be able to utter the pass phrase with speech characteristics similar to those of the client.

A fairly unsophisticated approach to spoofing a speaker-verified pass-phrase protected system is illustrated in Figure I and is termed herein as a *playback attack*. To execute a playback attack, an intruder obtains a recording of a client uttering his or her pass phrase while the client accesses the system. The intruder then phones the system, claims the identity of the true client, and plays back the recording of the client's pass phrase. The ready availability of inexpensive good quality recording devices reduces the challenge of executing such attacks, thus increasing their threat and the need to safeguard against them.

Although text-prompted systems (*i.e.*, systems which prompt the user to utter a randomly selected phrase for each access attempt) can be used to eliminate the risk associated with playback attacks, they do this at the expense of sacrificing the protection afforded by knowledge of a client-specific pass phrase, thus making them more vulnerable to other types of attacks. Other vulnerabilities of speaker verification systems are discussed in [1], [2], [3]. The present work is focused on safeguarding a pass-phrase speaker-verfied system against playback attacks.

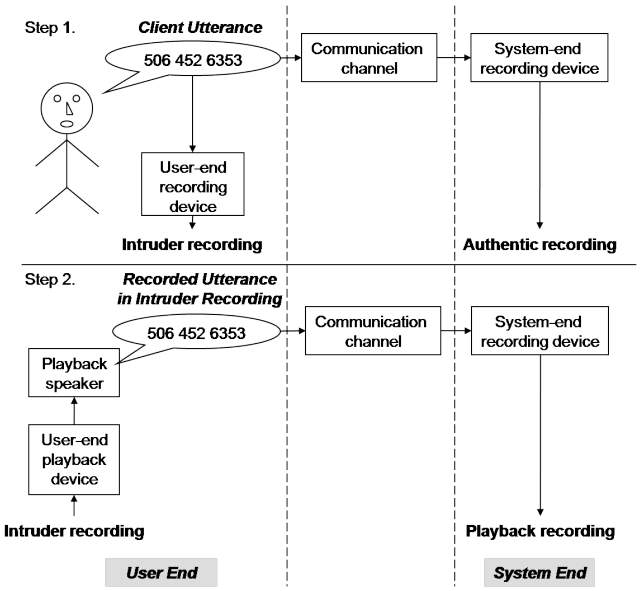



Fig. 1. Execution of a playback attack

In this paper, a methodology for detection of playback attacks is presented. The proposed methodology assumes that the utterance associated with the intruder's recording was produced by a client when accessing the system; it further assumes that each time a client accesses the system, the system-end waveform of the client's pass-phrase utterance is recorded, compressed, and stored by the system. The proposed approach exploits the random nature of human speech production which ensures the uniqueness of each utterance. For each system access, the incoming pass-phrase utterance is compared to the client's stored utterances (or representations there of). A playback attack is declared if the incoming utterance is deemed too similar to any of the stored utterances.

The design of the playback attack detector (PAD) will thus require: identification of a feature set capable of capturing

978-1-4244-1688-2/08/$25.00 

the utterance-specific characteristics of a speech waveform, a similarity measure to assess and quantify the similarity of two waveforms as represented by their features, and a decision rule which maps the resulting similarity scores into a playback-attack/no playback-attack decision.

The remainder of the paper is organized as follows. In section II, the challenges of the PAD task are reviewed. In section III and IV the design choices and algorithmic details of the proposed PAD are discussed. The data set and methodology used to assess the performance of the proposed PAD are described in section V and VI, respectively. Performance results are reported in section VII. Finally, in section VIII, conclusions are stated and future work is proposed.

## II. Challenges

Given the assumptions underlying the proposed methodology, the task of playback attack detection can be reduced to the task of determining whether a given pair of system-end waveforms originated from the same utterance or different utterances. The task is complicated by the fact that the system is assumed to be accessed remotely via telephone. This implies that the system-end waveforms are noisy channel-distorted versions of their associated user-end waveforms. Thus, even when there is a nondisputable association between the user-end waveforms associated with a client's original utterance and the playback of the intruder's recording, the variations in channel characteristics and noise from one transmission to the next will work in the intruder's favor to obscure the association between the resulting system-end waveforms.

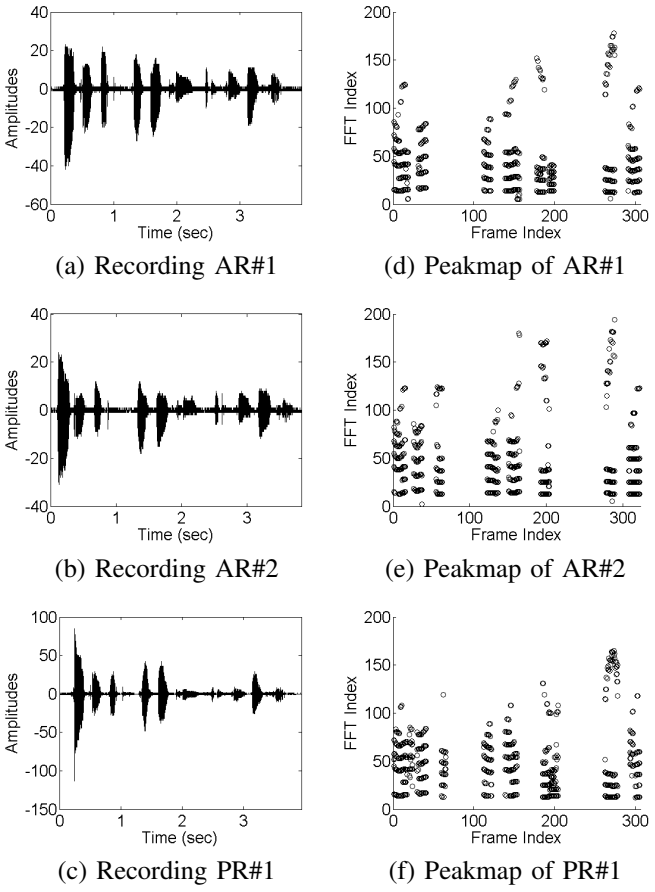
The task is also complicated by practical considerations such as storage requirements and processing time. In general, the more times a client accesses the system with a given pass phrase, the greater will be the storage requirements and processing time of the PAD. Thus a practical PAD design requires the identification of a parsimonious feature set capable of capturing the utterance-specific information needed to accomplish the goals of the PAD. It also requires a similarity measure, with low computational complexity, for the purpose of assessing the similarity of two speech waveforms as represented by their features.

## III. Fundamental design

Fundamental to a successful PAD design will be the identification of an appropriate feature set and similarity measure. The proposed feature set and similarity measure are discussed below.

### A. Proposed Feature Set

Considering the challenges and objectives of the PAD task, the ideal feature set should have the following characteristics: features should capture utterance-specific information (*i.e.*, feature values should vary significantly from one utterance to the next); features should be robust to channel noise and convolutional distortion (*i.e.*, the features of the user-end waveform should be very similar to those of the system-end waveform); and the feature set should lend itself to compact storage.




(a) Recording AR#1 (d) Peakmap of AR#1

(b) Recording AR#2 (e) Peakmap of AR#2

(c) Recording PR#1 (f) Peakmap of PR#1

Fig. 2. Feature sets of the authentic recordings and a playback recording. All recordings contain the phrase "506 452 6353", and they are made via landline by participant B.

The proposed feature set includes the indices of the *voiced* frames of the speech waveform and the FFT bin numbers of the $N_p$ highest peaks of the magnitude spectrum for each voiced frame. This feature set can be used to create a terse time-frequency plot of the waveform, termed herein as a *peakmap*. The peakmap has a representation as a sparse $N_t \times N_k$ binary-valued matrix, where $N_t$ denotes the number of frames into which the waveform is segmented and $N_k$ denotes the number of frequency bins (half the chosen FFT size). Peakmap elements are assigned a value of 1 if both the associated frame index corresponds to a voiced frame and the associated frequency bin index corresponds to one of the frame's $N_p$ highest spectral peaks; all other elements are assigned a value of 0.

Examples of speech waveforms and their associated peakmaps for the case when $N_p = 5$ are shown in Figure 2. The figure includes two authentic recordings (labeled as AR#1 and AR#2) and one playback recording (labeled as PR#1) which involves the same utterance as AR#1.

Note the proposed feature set captures utterance-specific in-

formation such as the duration of the various voiced/unvoiced segments and the trajectories of the high-energy harmonics within the voiced segments of speech. Furthermore, assuming the communication channels of interest to have relatively flat frequency responses over the frequency range spanned by the high-energy harmonics of the voiced speech segments, it is reasonable to expect that the frequency bin locations of the high energy spectral peaks will be unaffected by the transmission process.

The proposed feature set can be compactly stored. The precise storage requirements depend on the number of voiced frames and the number of unvoiced segments. A total of $N_p$ bytes are required to store the FFT indices for each voiced frame. Each unvoiced segment can be stored with two bytes, one byte is used to indicate an unvoiced segment, the other byte is used to indicate the number of frames in the segment. On average, a compression ratio greater than 50 is obtained when comparing the size of the 8 bits per sample, 8 kHz sampled WAV files storing the system-end recording to the size of the file storing the associated feature set.

### *B. Proposed Similarity Measure*

The proposed similarity measure operates on two peakmaps to produce a similarity score between 0 and 1 with higher scores indicating more similarity.

To find the similarity score, the cross correlation of the two peakmaps is evaluated solely as a function of the frame displacement variable, $\tau$; the frequency bin displacement variable is restricted to a value of zero. For a given value of $\tau$, the value of the cross-correlation is given by the number of 1s in the element-wise product of the two peakmaps when one peakmap is displaced by $\tau$ frames relative to the other. Normalization by the square root of the product of the number of ones in each of the two peakmaps yields the normalized cross correlation function. The maximum value of the normalized cross correlation function is used as the similarity score.

To illustrate how peakmaps vary from one utterance to the next, the peakmaps of the two authentic recordings, AR#1 and AR#2, shown in Figure 2, are cross correlated as described above. In Figure 3, the two peakmaps are then superimposed on the same plot with the frames of one being displaced relative to the frames of the other so as to illustrate the alignment between peakmaps which resulted in the maximum value of the cross correlation function. A similar comparison of the peakmaps of authentic recording AR#1 and playback recording PR#1 (both associated with the same utterance) is shown in Figure 4. In both figures, a zoom-in of the portion of the peakmaps enclosed by the dashed rectangle is provided. In both figures, the zoom-in portion corresponds to the first 3 in the pass phrase "506 452 6353". As would be expected, there is a better match between the peakmaps of Figure 4 than there is between the peakmaps of Figure 3. The comparison of the peakmaps for AR#1 and AR#2 yields a relatively low similarity score of 0.1875; whereas, the comparison of the peakmaps for AR#1 and PR#1 yields a much higher similarity score of 0.5223.

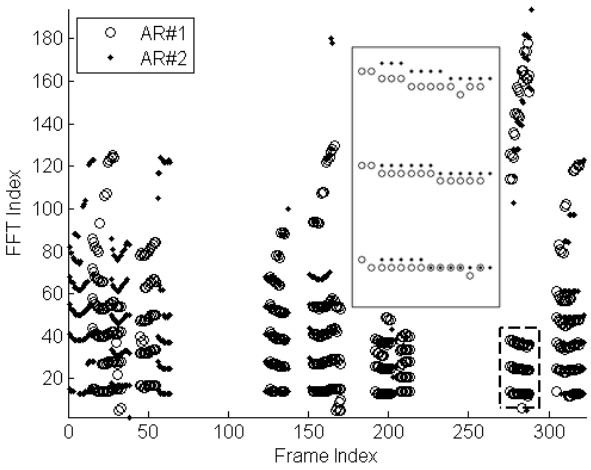



Fig. 3. Comparison of peakmaps for AR#1 and AR#2

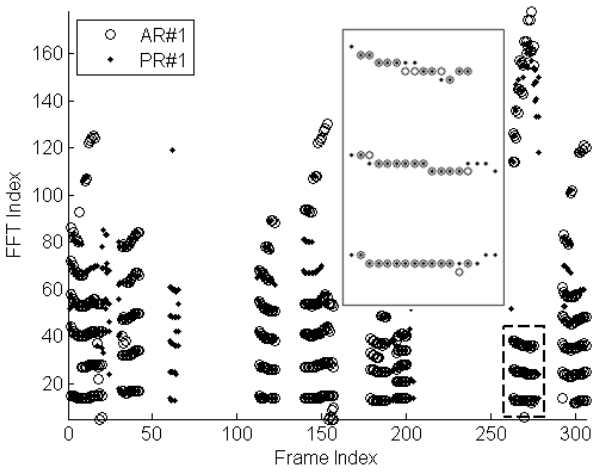



Fig. 4. Comparison of peakmaps for AR#1 and PR#1

It is also worth noting that the binary nature of the peakmap greatly reduces the computational complexity normally associated with computation of a cross correlation function. This is primarily due to the fact that the element-wise multiplication of two peakmaps can be replaced by the element-wise AND of the two binary-valued matrices. Note also that if the response time of the PAD becomes too long due to a large number of stored feature sets, the system could require the client to change his or her pass phrase.

## IV. ALGORITHMIC DETAILS

Each time a client requests access to the system, the PAD processes the *incoming recording* (*i.e.*, the system-end recording of the pass-phrase utterance) and makes a decision as to whether the incoming recording is an authentic or playback recording. The PAD algorithm consists of three stages: extraction of features from the incoming recording; comparison

of the incoming and stored feature sets; and classification of the incoming recording as either playback or authentic.

### A. Feature Extraction

The incoming recording (sampled at 8 kHz) is first divided into overlapping Hamming-windowed frames of length 256 samples using a frame interval of 80 samples. The 512 point FFT of each frame is then computed and the FFT indices of the five highest peaks of the one-sided magnitude spectrum are stored.

The next step of feature extraction is to assign voiced/unvoiced labels to each frame of the waveform. The method used in selecting voiced frames is based on the assumption that the harmonics associated with the $N_p = 5$ highest spectral peaks will remain fairly constant from frame to frame within the highly voiced segments of speech. Hence, within these segments, harmonic frequency tracks can be established by connecting peaks in one frame to their associated peaks at nearby frequencies in the neighboring frames. Following a procedure similar to frame-to-frame peak matching in [4], connections between peaks in neighboring frames are established where possible. Note that peaks in neighboring frames can only be connected if they are located within a few FFT bins of each other and that a peak in one frame can be connected to at most one peak in each of the two neighboring frames. Thus, assuming 5 peaks per frame, there will be a total of 10 connections that can be made to/from the peaks in a neighboring frame. Frames for which at least 8 of the 10 connections are made, are initially labeled as *voiced*; all remaining frames are labeled as *unvoiced*. After initial labels are assigned, groups of consecutive voiced frames are formed and the following two label refinements are made. If two groups of voiced frames are separated by a single unvoiced frame, the unvoiced frame is relabeled as *voiced*. Any group with fewer than five voiced frames is relabeled as *unvoiced*.

Once the voiced labels have been finalized, the resulting feature set can be expanded into a peakmap as described in section III-A

### B. Comparison with Stored Feature Sets

Assuming the claimed client has previously accessed the system a total of $N$ times using his/her current pass phrase, the system will have stored a total of $N$ feature sets representing the utterances associated with these previous system accesses. The similarity of the incoming feature set with each of the stored feature sets will be assessed using the similarity measure of Section III-B. This will yield a total of $N$ similarity scores, which can be used to label the incoming recording as an authentic recording or a playback recording.

### C. Classification of Incoming Recording

The decision as to whether the incoming recording is classified as an authentic recording or a playback recording is based on the maximum value of the $N$ similarity scores. The max similarity score is compared with a predetermined threshold. If the max score is greater than the threshold, indicating a high level of similarity between the incoming feature set and a stored feature set, the incoming recording will be classified as a playback recording; otherwise, it will be classified as an authentic recording.

## V. Database description

For the purpose of evaluating the proposed PAD's performance, a database containing intruder recordings, authentic recordings, and playback recordings was collected.

Over a span of four months, four participants each partook in 90 recording sessions. The participants were equally divided between the two genders and within each gender class, one of the participants was a native English speaker and the other was a non-native English speaker.

For each recording session, the participant connected to the system using one of three channel types (landline, cellular, or VoIP) and uttered the phrase "506 452 6353" while holding the channel-input microphone approximately one inch from the side of the mouth. The utterances were recorded at the system end (8 bits per sample and a sampling rate of 8 kHz) using a computer equipped with a telephony board. This resulted in a total of 90 authentic recordings for each participant (30 for each channel type). For 30 of the 90 recording sessions (10 sessions for each channel type), the participant's utterance was simultaneously recorded at the user end using a digital voice recorder (DVR), set at a sampling rate of 44.1 kHz. During these 30 sessions, the participant would hold two microphones (both the channel-input microphone and the DVR microphone) approximately one inch from and on opposite sides of the mouth.

For each of the 30 intruder recordings, three playback recordings were made (one for each channel type) resulting in a total of 90 playback recordings for each participant. Playback recordings are made at the system end in the same way as the authentic recordings; the difference occurs at the user end. When making playback recordings, the intruder recording is played back, at the user end, through a stereo speaker connected to the DVR; the channel-input microphone is held approximately one inch in front of the speaker.

The microphones used in the collection process were all electret.

## VI. PERFORMANCE EVALUATION

The performance of the proposed PAD is evaluated using the database described in Section V. For each participant, the 30 authentic recordings for which intruder recordings exist are used as *stored recordings*. The 60 remaining authentic recordings and 90 playback recordings are used as *incoming recordings*.

For each participant, the PAD methodology of Section IV is applied to each of the 150 incoming recordings. For each incoming recording, a max similarity score is found, representing the maximum similarity between the incoming recording and the 30 stored recordings. A scatter plot of the max similarity scores is shown for each of the four participants in Figure 5. Note that for each participant, the 90 scores for

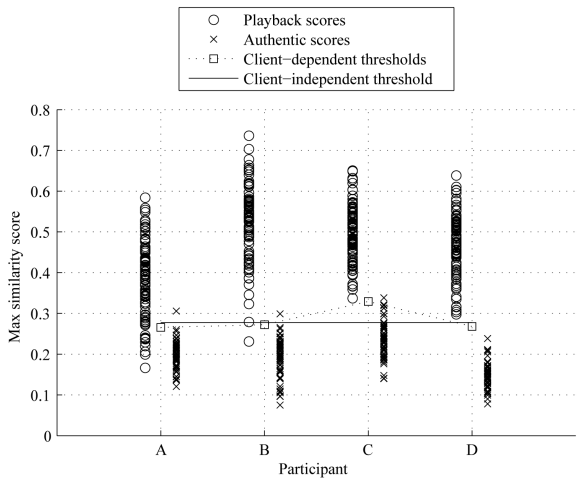



Fig. 5. Scatter plots of the similarity scores

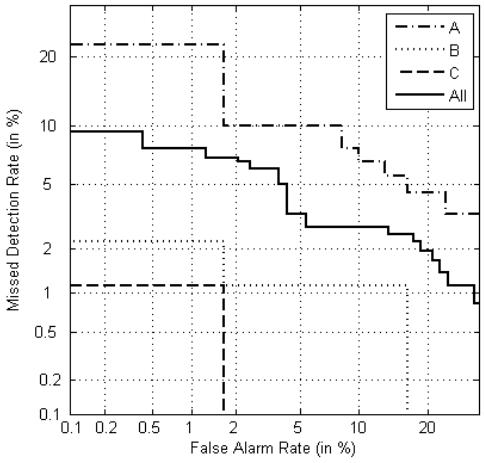



Fig. 6. DET curves

TABLE I
MISSED DETECTION RATES AND FALSE ALARM RATES

| Threshold | Client-dep. | | | | Client-indep. | | | |
|---|---|---|---|---|---|---|---|---|
| Participant | A | B | C | D | A | B | C | D |
| MDR | 10 | 1.1 | 0 | 0 | 12.2 | 1.1 | 0 | 0 |
| FAR | 1.7 | 1.7 | 1.7 | 0 | 1.7 | 1.7 | 13.3 | 0 |
| *in percentage (%). | | | | | | | | |

the playback recordings are plotted with an 'o' and offset to the left, whereas the 60 scores for the authentic recordings are plotted with an 'x' and offset to the right. In principle, we would like to see good separation between the playback scores and the authentic scores. From the plot, we see that playback scores are typically much higher than the authentic scores. With the exception of participant A, there is good separation between each participant's playback and authentic scores; several of the playback scores for participant A are down in the authentic score range. As a final remark, we note that although there is good separation between the authentic and playback scores for participant C, the authentic scores for participant C tend to be higher than the authentic scores of other participants.

Given a threshold value, the PAD labels the incoming recording as 'playback' if the max similarity score is greater than the threshold and 'authentic' otherwise. In general, the PAD can make two types of classification errors. A *false alarm* occurs when the PAD labels an authentic recording as 'playback'; in this case, the PAD incorrectly declares a playback attack and the client is denied access to the system. A *missed detection* occurs when the PAD labels a playback recording as 'authentic'; in this case, the PAD fails to detect a playback attack resulting in the possibility that the intruder will gain access to the system. It is worth noting that failure to detect a playback attack does not necessarily result in the intruder gaining access to the system. In particular, if the playback attack was not detected due to severe channel distortion, it is possible that the distortion will also prevent the speaker verification system from verifying that the pass phrase was spoken by the claimed client.

In general, as the PAD threshold is increased from 0 to 1, the Missed Detection rate (MDR) will increase from 0% to 100% while the False Alarm rate (FAR) will decrease from 100% to 0%. This results in a tradeoff between the FAR and the MDR; the optimal threshold selection will depend on the cost of a false alarm versus the cost of a missed detection. A DET curve [5] is a convenient way of showing the various tradeoffs (*i.e.*, the MDR/FAR performances) that can be achieved with different thresholds. Figure 6 shows both the client-dependent and client-independent DET curves. Similar to the scatter plot, we see that participant A has the worst performance with an equal error rate (EER) of approximately 8%; whereas all other participants have an EER less than 2% with participant D having an EER of 0% (thus the absence of the DET curve for D). When using a client independent threshold, the resulting overall EER is 4.2%.

Superimposed on the scatter plot of Figure 5 are the client-dependent and client-independent thresholds chosen so as to minimize the overall number of errors. The missed detection rates and false alarm rates that result from these thresholds are summarized in Table I. As seen from the table, the use of a client-dependent threshold as opposed to a client-independent threshold serves to significantly reduce the false alarm rate for client C; otherwise the results of the two threshold types are comparable.

## VII. DISCUSSION

In an attempt to understand the relative poor performance of participant A, peakmaps associated with several of the playback/stored recording pairs that yielded low similarity scores were investigated. Two problems were identified.

The first problem that was the relatively low pitch ($\approx 90$ Hz) of participant A which resulted in the low harmonic frequencies falling below the pass band of the channel frequency

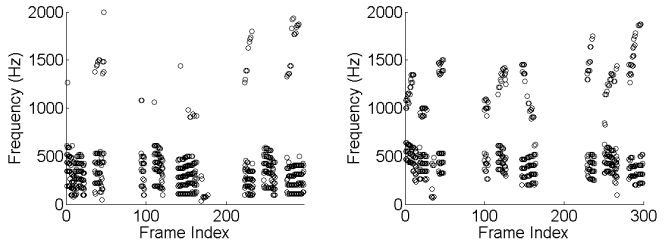



Fig. 7. Elimination of lower frequency peaks, a) peakmap of an intruder rec. and b) peakmap of an authentic rec.

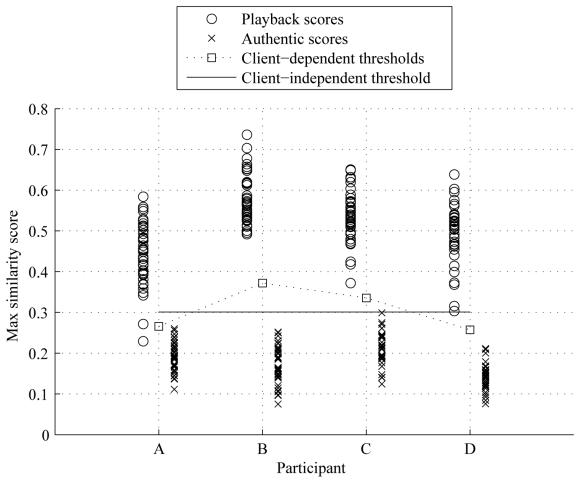



Fig. 8. Scatter plot of the similarity scores for the recordings made via landline and cellular line

response (Figure 7). These peaks were thus attenuated by the channel, making them less likely to be included in the peakmap. In addition, the replacement peaks tended to be less significant than the unattenuated lower harmonics would have been which resulted in their frequency locations being more easily affected by channel distortion and noise.

The second problem was a result of temporal distortion caused by the VoIP channel. Examination of one of the VoIP stored recordings for participant A revealed that, in comparison to the associated intruder and playback recordings, the last part of the signal was delayed relative to the first part. This was most likely a result of the way in which delayed packets are handled by the VoIP service provider. The fact that the stored recording was distorted resulted in low scores for all three of the associated playback recordings.

In order to get a rough idea as to how the overall PAD performance was affected by the VoIP channels, all stored and incoming recordings that were made via a VoIP channel were eliminated. Figure 8 shows the scatter plot of the resulting max scores. Comparison with the scatter plot of Figure 5, reveals that many of the lower playback scores from other participants were also associated with the VoIP channel.

## VIII. Conclusion and future work

In this paper, a PAD that can be used in guarding speaker verification systems against playback attacks has been proposed and evaluated. Fundamental to the PAD design was the identification of a set of features which varies significantly from one utterance to the next while also demonstrating some resistance to channel noise and distortion. Utterance-specific information captured by the feature set includes the duration of the various voiced/unvoiced segments and the trajectories of the high-energy harmonics within the voiced segments of speech. The similarity of two utterances is assessed by means of a proposed similarity measure which operates on the associated feature sets.

The PAD was evaluated on data from four participants and three communication channel types. Overall, the scores assigned by the PAD to the playback recordings were higher than the scores assigned to the authentic recordings, thus allowing the PAD to successfully detect most playback attacks with a relatively low false alarm rate.

The performance evaluation identified two weaknesses of the proposed PAD: the first was that the features extracted from utterances of low-pitch clients are less robust to channel distortion than those from higher-pitch clients; the second was the sensitivity of the feature set to the temporal distortion which is occasionally caused by the VoIP channels.

Future work will investigate solutions to these two weaknesses. One possibility for dealing with the temporal distortion introduced in VoIP channels is to assess similarity scores for each voiced segment. An overall similarity score could then be obtained as a weighted combination of the various segment scores. Possible solutions for low-pitch clients include limiting the peak selection to a frequency range within the pass band of the communication channels and/or reducing the number of selected peaks per voiced frame.

Future work will also include the evaluation of the proposed PAD on a database which includes a variety of different pass phrases.

## Acknowledgment

This work was supported by the Natural Science and Engineering Research Council of Canada.

## References

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

[2] Y.W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, Oct, 2004, pp. 145–148.

[3] D. Matrouf, J.F. Bonastre, and C. Fredouille, "Effect of speech transformation on impostor acceptance," in *ICASSP*, 2006, pp. 933–936.

[4] Robert J. McAulay and Thomas F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 4, pp. 744–754, August 1986.

[5] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech '97*, Rhodes, Greece, 1997, pp. 1895–1898.