SCORE NORMALIZATION IN PLAYBACK ATTACK DETECTION

Wei Shang and Maryhelen Stevenson

University of New Brunswick
Department of Electrical and Computer Engineering
P.O. Box 4400, Fredericton, New Brunswick, Canada E3B 5A3

ABSTRACT

The task of a playback attack detector (PAD) is to decide whether an incoming recording shares the same originating utterance as any of N stored recordings. All recordings are noisy channel-distorted versions of the same phrase uttered by the same person; the originating utterances of the N stored recordings are assumed to be distinct. The proposed approach makes a decision based on a set of N similarity scores which quantify the similarity between the incoming recording and each of the N stored recordings. Although satisfactory results are obtained by thresholding the maximum of the N scores using speaker and phrase (SaP)-dependent thresholds, it is shown that the use of a relative similarity score (a normalized version of the maximum similarity score) results in significant performance improvements especially in the case when the incoming recording is a severely distorted version of a stored recording utterance, as well as for the case when SaP-independent thresholds are used.

Index Terms— Score normalization, test normalization, playback attack detection, speaker verification

1. INTRODUCTION

Because of the robustness and convenience it offers, speaker verification ([1, 2]) has become a popular technique used in verifying peoples' identities. A commonly made assumption in speaker verification is that the speech utterance is spoken by the person who is interacting with the system; this is not true in the case of a playback attack. In a *playback attack* (see Figure 1), an intruder obtains a user-end recording of a client uttering their pass phrase while accessing the system and later plays back the recorded utterance in an attempt to gain unauthorized access to the system. The various recordings shown in Figure 1 are defined as follows: the *intruder recording* is a user-end recording of the original utterance; the *authentic recording* is a system-end recording of the channel-distorted original utterance; and the *playback recording* is a system-end recording of a channel-distorted playback version of an intruder recording.

Because playback attacks are effective and easy to implement, they pose a serious threat to the operation of a speaker-verified pass phrase protected system. Although text-prompted systems can be used to eliminate the risk associated with playback attacks, they do so at the expense of sacrificing the protection afforded by the knowledge of a speaker-specific pass phrase. To offer some protection against playback attacks, a playback attack detector (PAD) has been developed[3]. The PAD capitalizes on the hypothesis that, due to the random nature of the speech production process, distinctions can always be made between two utterances of the same phrase by the same person. Thus, if little distinction is found between two recordings, it is likely they originate from the same utterance.

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

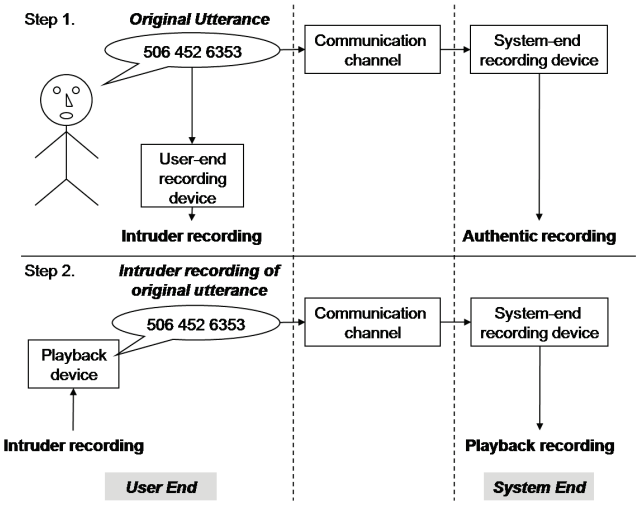


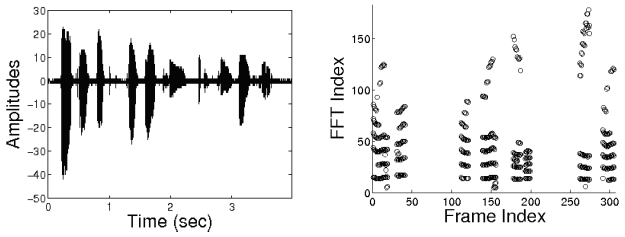
Fig. 1. Execution of a playback attack

While the PAD has demonstrated satisfactory performance, results from a preliminary study [4] revealed that various factors, *e.g.*, speaker, pass phrase, communication channel, playback device and number of stored recordings, can affect the PAD score densities. Such effects limit the performance achievable by a PAD using a global threshold and suggest the use of factor-dependent thresholds and/or score normalization techniques. In this paper, we propose the use of a relative similarity score as a means of reducing such effects and thus improving PAD performance.

The remainder of the paper is organized as follows. In Section 2, a brief description of the PAD algorithm is provided, and the relative similarity score is introduced as a replacement for the maximum similarity score in the PAD's decision making process. In Section 3, the database used in this study is described and the experimental procedures are explained. In Section 4, results from the experiments are presented and comparisons are conducted between the performances of the original and the new scoring methods. Finally, in Section 5, conclusions are made, and future work is described.

2. THE PAD ALGORITHM

The PAD algorithm consists of three stages, namely, *feature extraction*, *similarity measure*, and *attack/non-attack classification*. Operations at each stage are described briefly below; implementation detail is provided in [3]. Two authentic recordings, labeled as AR#1



(a) time-domain waveform (b) peakmap of signal in (a)

Fig. 2. Time-domain waveform of an utterance and its peakmap

and AR#2, and a playback recording (associated with AR#1), labeled as PR#1, are used to demonstrate the operation of the PAD.

2.1. Feature Extraction

The PAD represents each recording in terms of a *peakmap*, which consists of time and frequency locations of the five highest spectral peaks in each of the voiced frames. The incoming recording (sampled at 8000 Hz) is first divided into overlapping Hamming-windowed frames with a frame size of 32 ms (or 256 samples) and a frame interval of 10 ms (80 samples). The spectrum of each frame is then found via a 512-point FFT. In essence, a peakmap is a 2-D binary matrix of size N_t by N_f, where N_t is the total number of frames in the utterance and N_f is the number of FFT bins. Element (i, j) of the peakmap is assigned a value of 1 if frame j is voiced and one of the five highest spectral peaks for frame j is located in the i^{th} FFT bin; otherwise it is assigned a value of 0.

Figure 2 depicts both the time-domain waveform and the peakmap of the utterance stored in AR#1. Notice that the time and frequency locations where spectral peaks exist are marked with an "o".

2.2. Similarity Measure

The similarity between two peakmaps is quantified by 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 τ; the frequency bin displacement is restricted to a value of zero. For a given value of τ, 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 τ frames relative to the other. The similarity score is obtained by normalizing the maximum value of the cross-correlation function by the square root of the product of the number of ones in each of the two peakmaps.

Figure 3 depicts the peakmaps 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 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 (corresponding to the first 3 in the pass phrase "506 452 6353") is provided. As would be expected, there is a better match between the peakmaps in Figure 4 than those 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.

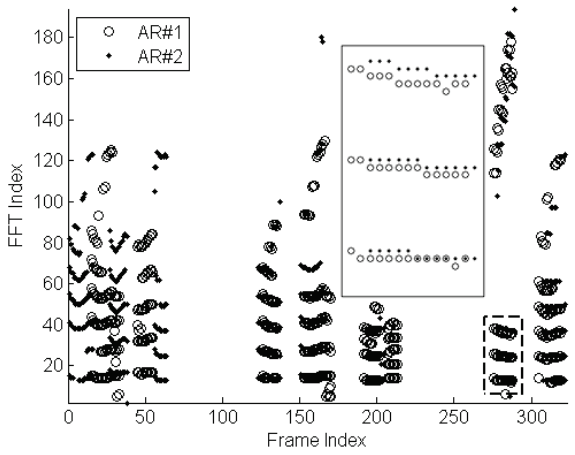


Fig. 3. Comparison of two authentic recordings

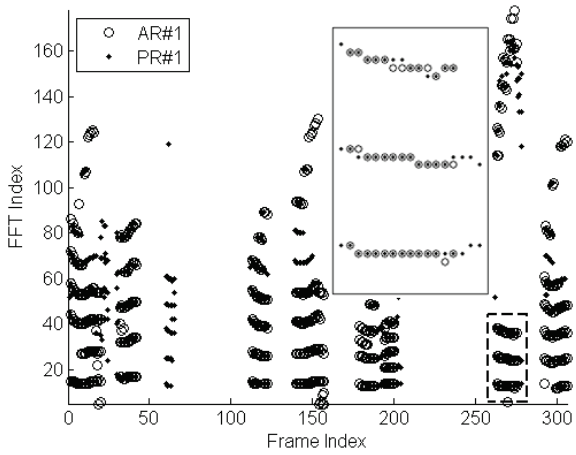


Fig. 4. Comparison of a playback recording and an authentic rec.

2.3. Attack/non-attack Classification

The PAD must label each incoming recording as either an authentic or a playback recording. Assuming the claimed speaker has previously accessed the system a total of N times using his current pass phrase, the system will have stored a total of N peakmaps representing the utterances associated with these previous system accesses. The similarity of the incoming peakmap with each of the stored peakmaps is then assessed yielding a set, S_N, of N similarity scores.

In previous work, the PAD's labelling decision was based on the *maximum similarity score*, $s_{\max} \equiv \max(S_N)$. If $s_{\max}$ is greater than the threshold, t_m, indicating a high level of similarity between the incoming peakmap and a stored peakmap, the incoming recording will be labelled as a playback recording; otherwise, it will be labelled as an authentic recording.

The newly proposed labelling decision is based on the *relative similarity score*, $s_{\rm rel}$, a normalized version of $s_{\max}$, which is calculated as:

$$s_{\rm rel} = \frac{s_{\max} - \mu_{N-1}}{\sigma_{N-1}} \quad (1)$$

where μ_{N-1} and σ_{N-1} denote the sample mean and standard deviation of the $N-1$ scores in S_{N-1} and where S_{N-1} is the set of scores which result from discarding $s_{\max}$ from the set S_N. Similar to the test normalization, introduced in [5] for speaker verification, we note that the normalizing parameters, μ_{N-1} and σ_{N-1}, will change from one incoming recording to the next.

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 speaker 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.

3. DATABASE AND EXPERIMENTAL PROCEDURE

A database containing a total of 4320 recordings is used in evaluating the PAD's performance. Depending on the speaker and pass phrase involved, the database is equally divided into 12 sets with each of the 12 sets representing a particular speaker and phrase (SaP) combination; there are a total of four speakers and three pass phrases. For each SaP combination, there are 360 recordings, 30 of which are stored recordings and the remaining 330 are incoming recordings. Of the 330 incoming recordings, 60 are authentic recordings and the remaining 270 recordings are playback recordings.

Three communication channels and three playback devices were used in making the playback recordings. Thus, for each stored recording, there are nine corresponding playback recordings that were made from one of the nine channel/playback device combinations. Of the 270 playback recordings, 90 of them are made with a poor quality playback device. These recordings are NOT included in the experiments unless otherwise specified.

For each of the 12 SaP sets, the experiment is conducted as follows: one by one, each of the 240 incoming recordings (180 playback recordings and 60 authentic recordings) is compared to each of the 30 recordings in the stored SaP set resulting in a set of 30 similarity scores for each incoming recording. As described in the previous section, the two scores, $s_{\max}$ and s_{rel}, are then determined for each incoming recording. Subsequently, the SaP-dependent equal error rate threshold is determined for each SaP set and the resulting *posterior* error rates for each SaP set are recorded.

The scores from the 12 SaP sets are then combined, and a SaP-independent threshold that yields the equal error rate for the combined set of scores is found. Using the SaP-independent threshold for the combined score set, the overall error rate as well as the SaP-specific missed detection rates (MDR) and false alarm rates (FAR) are determined.

4. RESULTS AND DISCUSSION

Figure 5 depicts the scatter plots of the playback maximum similarity scores $S^p_{\max}$ and the authentic maximum similarity scores $S^a_{\max}$ obtained from each of the 12 SaP sets. For each SaP set, $S^p_{\max}$ are plotted with lighter dots and offset to the left, whereas $S^a_{\max}$ are plotted with darker dots and offset to the right. Similarly, scatter plots for playback relative similarity scores S^p_{rel} and authentic relative similarity scores S^a_{rel} are plotted in Figure 6. Notice that in both figures, and for all SaP sets, the mean of the playback scores is clearly distinguished from the mean of the authentic scores. Also notice, from Figure 6, that the range of S^p_{rel} is much wider than that of S^a_{rel}.

Table 1 lists the *posterior* error rates that result for each SaP set when the SaP-dependent thresholds are applied. Note that since

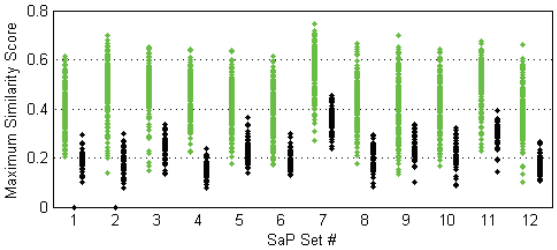


Fig. 5. Scatter plots of $S^p_{\max}$ and $S^a_{\max}$ from 12 SaP sets

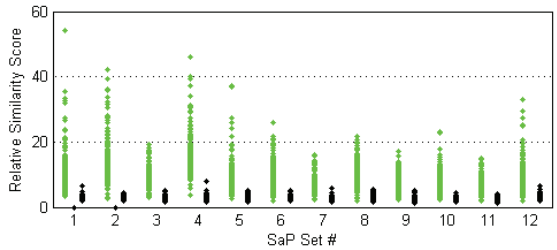


Fig. 6. Scatter plots of S^p_{rel} and S^a_{rel} from 12 SaP sets

equal error rate thresholds are used, the SaP-specific missed detection rates and false alarm rates are given by the SaP-specific error rates in the table. From the table, we observe that when s_{rel} is used instead of $s_{\max}$, the error rates are lower for 9 out of the 12 SaP sets. Significant reductions, sometimes greater than 50%, are seen for some SaP sets.

When using SaP-independent thresholds, the overall error rate drops from 11.94% when using $s_{\max}$ to 6.81% when using s_{rel}. Table 2 lists the SaP-specific MDRs and FARs for the 12 SaP sets when the SaP-independent threshold is used. Notice that the error rates resulting from $s_{\max}$ can be as high as 73.33%, whereas the highest error rate when using s_{rel} is 21.67%. This observation indicates that the use of a single threshold could cause significant performance degradation in some SaP sets if $s_{\max}$ is used. In contrast, because s_{rel} aligns the authentic relative similarity scores from the various SaP sets, the performance degradation resulting from the switch from SaP-dependent thresholds to SaP-independent thresholds is much less.

Similar observations can be made in Figure 5 and Figure 6. Notice in Figure 5 that while there are noticeable separations between $S^p_{\max}$ and $S^a_{\max}$ for each SaP set, the means of $S^p_{\max}$ and $S^a_{\max}$ vary from one SaP set to the next. Consequently, during the deployment of the PAD, the use of a SaP-independent threshold is unlikely pro-

Table 1. Overall error rates (%) for each SaP set obtained with SaP-dependent thresholds

SaP set	#1	#2	#3	#4
$s_{\max}$	5.00	4.58	6.25	1.25
s_{rel}	3.33	2.08	3.33	1.67
SaP set	#5	#6	#7	#8
$s_{\max}$	11.67	9.58	10.00	5.42
s_{rel}	6.67	6.67	8.33	8.33
SaP set	#9	#10	#11	#12
$s_{\max}$	18.33	10.00	9.58	6.25
s_{rel}	9.58	5.00	5.42	9.58

Table 2. MDR (%) \FAR (%) for each SaP set when using SaP-independent EER threshold

Set	#1	#2	#3	#4
$s_{\max}$	22.78\0	7.78\0	5.56\8.33	5 \0
s_{rel}	1.67\6.67	1.67\3.33	3.33\3.33	0.56\8.33
Set	#5	#6	#7	#8
$s_{\max}$	13.89\8.33	18.89\0	0.56\73.33	8.89\0
s_{rel}	5.56\10	7.22\3.33	15.56\5	4.44\21.67
Set	#9	#10	#11	#12
$s_{\max}$	21.67\10	15.56\3.33	6.11\40	16.67\0
s_{rel}	14.44\5	8.33\3.33	12.78\0	6.11\11.67

duce satisfactory performance for all SaP sets. In contrast, the means of the various S^a_{rel} score sets (shown in Figure 6) are more or less aligned. As a result of the alignment, a SaP-independent threshold can be used to provide ease of implementation without causing much drastic performance degradation.

The s_{rel} method yields lower overall error rates than the $s_{\max}$ method does because under certain circumstances, the s_{rel} method can still correctly classify the incoming recordings that the $s_{\max}$ method fails to. For example, the $s_{\max}$ of a severely distorted playback recording might be well below the threshold t_m; however, the associated s_{rel} could still be higher than the threshold t_r provided that $s_{\max}$ is much higher than μ_{N-1}. In other words, even though the playback recording is very different from its associated stored recording, the fact that it is even more different from other stored recordings increases its likelihood of being a playback recording. For the purpose of demonstration, experiments with the playback recordings containing only the 90 recordings (per SaP set) that were made with a poor quality playback device are conducted. Figure 7 and 8 depict the corresponding scatter plots for the $s_{\max}$ and s_{rel} scores respectively. Notice that when $s_{\max}$ is used, there is significant overlap between the authentic scores and playback scores, whereas when s_{rel} is used, much less overlap is observed. Consequently, when s_{rel} is used instead of $s_{\max}$, the overall error rate resulted from the equal error rate threshold for the combined SaP set is reduced to 19.44% from 53.06%.

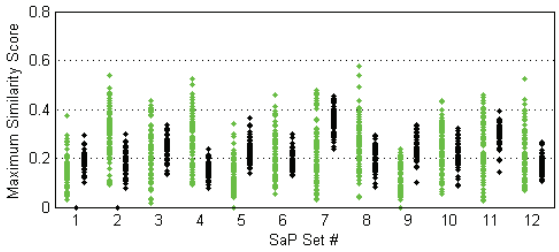


Fig. 7. Scatter plots of $S^p_{\max}$ and $S^a_{\max}$ with playback recordings made from a poor quality playback device

Some performance degradations are observed for SaP sets #4, #8 and #12 in Table 1. Note that the speaker is the same for these three SaP sets. The cause for the degradation is thought to be the fact that several of this speaker's authentic recordings are quite different from most of the stored recordings. Consequently, most of the similarity scores in S_N are located very close to 0, resulting in low values for σ_{N-1}, which in turn, results in high relative similarity score for these authentic recordings. A simple solution for this problem is to set a minimum allowable value for σ_{N-1} in the computation for s_{rel}.

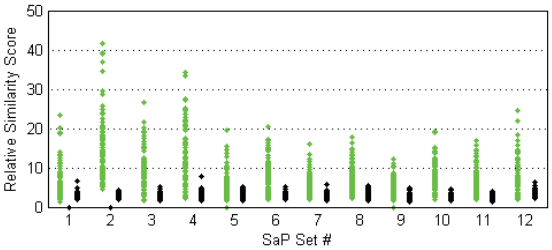


Fig. 8. Scatter plots of S^p_{rel} and S^a_{rel} with playback recordings made from a poor quality playback device

When the minimum allowable value is set at 0.05, the resulting overall error rate for the SaP-independent threshold is further reduced to 5.56% from 6.81%.

5. CONCLUSION AND FUTURE WORK

In this paper, the relative similarity score method is proposed as a score normalization scheme for playback attack detection; it is similar to the test normalization method used in speaker verification. Results show that the proposed method yields lower overall error rates for most SaP sets used in the experiments, with significant error rate reductions seen in many SaP sets. Moreover, the overall error rate obtained from the SaP-independent EER threshold is reduced from 11.94% when using $s_{\max}$ to 6.81% when using s_{rel}. When using the set of severely distorted playback recordings, application of an SaP-independent threshold to the set of $s_{\max}$ scores was unsuccessful, producing an overall error rate of 53.06%; in contrast, the s_{rel} scores resulted in an improved overall error rate of 19.44%.

Future work will determine how the mean and standard deviation of s_{rel} scores for authentic recordings change as the number of stored recordings increases, it is expected that this knowledge will provide insight regarding how the threshold should be adjusted as the number of stored recordings increases. Methods that can be used to find the optimal value for the minimum allowable value of σ_{N-1} in the computation of s_{rel} will also be investigated.

6. REFERENCES

[1] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, Magrin I. Chagnolleau, S. Meignier, T. Merlin, Petrovska D. Delacretaz, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430–451, 2004.

[2] J. P. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[3] Wei Shang and M. Stevenson, "A playback attack detector for speaker verification systems," in *Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on*, March 2008, pp. 1144–1149.

[4] W. Shang and M. Stevenson, "A preliminary study of factors affecting the performance of a playback attack detector," in *Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on*, May 2008, pp. 000459–000464.

[5] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42 – 54, 2000.