Using timing information in speaker verification

Charl J. van Heerden, Etienne Barnard

Human Language Technologies Research Group,
University of Pretoria / Meraka Institute, Pretoria, South Africa

Abstract

This paper presents an analysis of temporal information as a feature for use in speaker verification systems. The relevance of temporal information in a speaker's utterances is investigated, both with regard to improving the robustness of modern speaker verification systems and to detecting and deflecting recording attacks. It is shown that the use of timing information provides useful additional information that can be used to enhance the performance of verification systems, and that intra-speaker variability of typical tokens is sufficient (in comparison with typical noise-induced variability) to support the detection of recordings.

1. Introduction

1.1. Use of speaker identification and verification

Speaker verification systems are widely used to provide multilevel access control and prevent unauthorized use of computer and communication systems [1]. Although speaker verification systems are not completely secure, they are a powerful deterrent to fraud in combination with other security measures such as pin numbers, SIM cards or passwords.

Speaker identification is a related application of similar nature: in verification, a speaker claims to be someone and the system must verify this claim, while identification comprises the system to choose an individual from a database of speakers to select who is speaking. The issues addressed in this paper apply equally to both applications, and for simplicity we will focus mostly on verification.

1.2. Contexts for speaker verification

Speaker verification is generally applied in three different ways, depending on the text that a user is required to utter; these are, respectively, known as "text-independent", "text-dependent" and "text-prompted" systems. Each of these contexts is useful in its own right. The main advantage of text independent approaches is that the identity of a speaker can be verified in the background while the speaker is performing some spoken task, without the speaker even being aware of the verification process.

Text-dependent systems are generally more secure than text-independent systems, since the user can only say text that is known to the system beforehand, e.g. a password, name or telephone number. It is thus secure in the sense that the system can make a very accurate comparison between an existing template and the speech signal, and straightforwardly combines knowledge of a password with voice characteristics in performing verification.

Text-prompted systems are the most secure, since the system decides exactly what phrase should be said by the user. This makes it difficult to attack the system by playing recordings since the impostor will not know what text will be prompted beforehand (thought the widespread availability of digital recording devices reduces the effectiveness of this strategy - see below). On the other hand, this requires enrollment of all the phrases that may be prompted by the system, which users find tedious.

1.3. The problems of recordings; possible solutions

Biometric systems are considered the most secure access control techniques [2] available today. Speech is a biometric that, used in conjunction with traditional security systems, can greatly enhance access control when used in speaker verification systems. One of the specific weaknesses of speaker verification, though, is its susceptibility to attack by a recording of a person's voice being played.

To counter these attacks, text-dependent and text-prompted systems have been employed to make it either very difficult to obtain the required utterance (e.g. password in text-dependent system) or to obtain the correct sequence of prompts (e.g. random digits in text-prompted system). Both these techniques have been found to reduce the frequency of successful attacks, but with the advent of modern speech processing capabilities, even these techniques have reduced value. With modern technology, it is a relatively simple matter to synthesize whatever is prompted in a speaker's voice, if appropriate recordings had been made beforehand.

1.4. Temporal information for speaker verification and detection of recordings

Two types of errors can occur while employing speaker verification access control. These are false acceptance and false rejection errors. A tradeoff has to be made between the two, since overlap between different speakers'

models is inevitable.

Models that are typically used include statistical models when employing Hidden Markov Models (HMMs), template models when using dynamic time warping (DTW) and codebook models when vector quantization (VQ) is used [1]. Ideally, the speaker models must be unique. In practice, one can employ more distinguishing features to reduce the overlap between models. Current features that are used are (amongst others) linear predictor coefficients and mel frequency cepstral coefficients. However, it is likely that speakers also differ significantly in the duration assigned to segments they speak. As far as could be determined, such temporal information has not yet been incorporated as an independent or even complementing feature in speaker verification. One aim of the current research is therefore to determine whether temporal information can be used to improve the accuracy of modern speaker verification systems.

Another important application of temporal information is to detect when recordings or computer synthesized voices are being used in an attempt to gain unauthorized access to a system. The theory behind this is that a speaker's temporal information will be subject to substantial random variation. The probability of it being identical during different instances of normal speech is therefore very low, while a recording played twice or a computer synthesized voice should produce very similar detected times. Background and channel noise (amongst other factors) can produce some deviation in observed durations of e.g. a specific phoneme. To successfully detect recordings or synthesized voices, it is therefore necessary that this noise-induced variability must be substantially less than the real variability produced by speakers. If this is the case, as in figure 1 (the highlighted areas represent the acceptable area, where different instances of the same phoneme are expected to occur), acceptable recognition of recordings will be possible. Another aim of the present research is thus to determine whether the statistics of intra-speaker variation are sufficiently variable in comparison with noise-induced variability to detect recordings robustly.

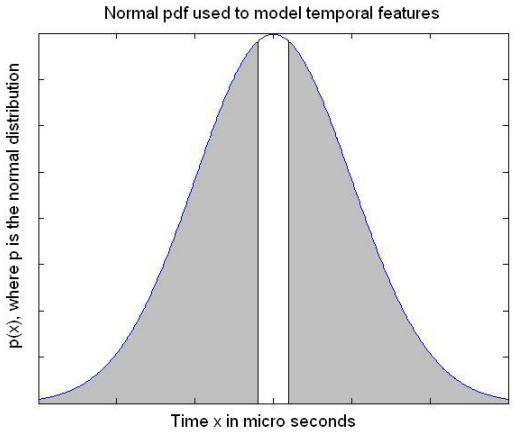

Figure 1: Dark areas show acceptable areas where normal variance in temporal information will cause recurring instances of the same phoneme to occur. The distribution is centered around a duration that occurs in an utterance which is potentially the source of a recording

2. Approach

2.1. Data collection

Speech was recorded with a desktop microphone from 26 people (8 female, 18 male) at 16kHz, 32 bits/sample. The speaker verification system used to test the data was constructed using HTK 3.2.1. The system is based on combination-lock-phrases [3], thus random combinations of the phrases "twenty-one" to "ninety-nine" were prompted to the user. The system recognizes these phrases using context-dependent triphone models, with mel-frequency cepstral coefficients (MFCCs) as input features. HMMs were generated for all triphones, with one Gaussian mixture per state, and a restricted grammar consisting of all combination-lock-phrases from "twenty-one" to "ninety-nine" (with multiples of ten omitted) was employed. This allowed for the creation of an efficient text-prompted system. The consequent word-recognition accuracy was above 90%, and formed the basis for the speaker-verification results presented below. To create speaker specific models, adaptation - first using maximum likelihood linear regression (MLLR) and subsequently maximum a-posteriori (MAP) adaptation [4] - was employed.

2.2. Baseline system

The initial speaker verification system created is similar to conventional systems, using acoustic likelihoods as the main speaker feature. We compared two approaches. First, a forced alignment was done, where the utterances were applied to the correct path through the phoneme network (as prompted) and scored based on their acoustic likelihoods. Secondly, a Viterbi word recognizer was used to determine the best fit to an utterance, using first the claimed speaker model and subsequently a universal background model. The associated frame log likelihood probabilities were then used to determine a normalized score for the speaker.

The frame log-likelihood probability ratios for the first nine speakers are shown in tables 1 and 2. (These were obtained on the test sets of the first and second recording sessions, respectively, based on training data from both sessions.) The scores in a given row reflect the average ratios obtained for speech by a particular speaker; the values in the corresponding column were obtained using the model of that speaker. The best (lowest) score in

every column is highlighted. As can be seen, the diagonal values are the smallest values in each column in all but one case, indicating that successful speaker identification was achieved using acoustic likelihood scores.

	1001	1006	1007	1008	1009	10010	10011
1001	**0.876**	0.996	1.03	0.991	0.95	0.95	0.959
1006	0.999	**0.901**	0.946	0.969	0.981	0.953	0.972
1007	1.012	0.940	**0.892**	0.992	0.993	0.976	0.987
1008	1.024	0.984	1.016	**0.915**	0.982	0.976	0.996
1009	0.983	1.005	1.039	0.984	**0.899**	0.962	0.958
10010	0.972	0.974	1.009	0.959	0.938	**0.881**	0.922
10011	0.958	0.979	0.994	0.975	0.913	0.909	**0.874**

Table 1: Acoustic frame log likelihood probabilities for speakers during session 1.

	1001	1006	1007	1008	1009	10010	10011
1001	0.957	1.079	1.046	1.017	1.018	1.018	0.987
1006	0.947	**0.882**	0.916	0.93	0.94	0.929	0.943
1007	0.956	0.936	**0.857**	0.945	0.968	0.956	0.964
1008	0.951	1.009	0.962	**0.879**	0.927	0.944	0.937
1009	0.904	0.973	0.948	0.91	**0.86**	0.906	0.895
10010	**0.878**	0.943	0.923	0.892	0.88	**0.846**	**0.874**
10011	0.924	1.004	0.973	0.94	0.92	0.919	0.881

Table 2: Acoustic frame log likelihood probabilities for speakers during session 2.

As can be seen in table 1 and 2, some scores from possible impostors are very close to those of the true speaker. Thus, the impostor may be falsely accepted by the system since the threshold value used as a cutoff for accepting or rejecting a speaker based on the score may not compensate for such a small difference. In the next section, we investigate whether this difference can be improved by incorporating temporal information.

3. Results

In this section, we report on various experiments that were performed to assess the utility of temporal information to improve on the baseline verification accuracy, and to detect possible recorded utterances.

3.1. Variability of times, within and across speakers

To understand the relationship between intra-speaker and inter-speaker variability in phone duration, a number of measurements were made on triphones as recognized by the speaker-independent models. Examples of the results obtained are shown in Figures 2, 3 and 4. These depict Gaussian distributions that were fit to the durations of different phonemes as spoken by a particular speaker across two different sessions, compared with the distributions obtained when all speakers' durations are pooled. A range of behaviours are observed: for some speakers and phonemes (as in Fig. 2), the intra-speaker differences are small and consistent across sessions, which is encouraging for our purposes. However, other speakers (see Fig. 3) or other phonemes (Fig. 4) produced less encouraging distributions.

The relative prevalences of these behaviours can be studied by developing a speaker-identification system similar to the one used in Tables 1 and 2, but with temporal information as distinguishing feature. The results, shown in Tables 3 and 4, are encouraging: the diagonal values are again generally the lowest for each speaker, thus suggesting that temporal features are potentially useful for both speaker identification and speaker verification.

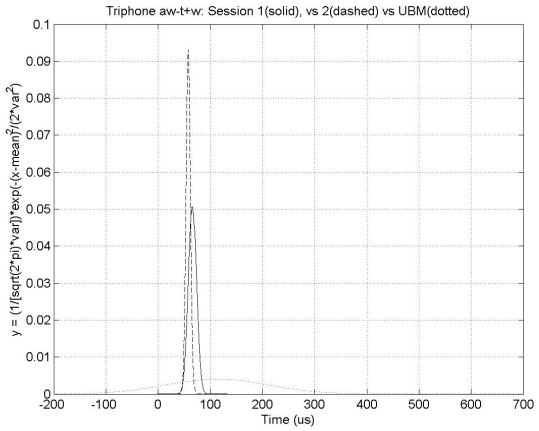

Figure 2: Speaker 1001 intersession variability for triphone aw-t+w in two different sessions, compared with the inter-speaker variability.

	1001	1006	1007	1008	1009	10010	10011
1001	**0.761**	0.8978	0.8956	0.8658	0.8565	0.8782	0.8982
1006	1.0085	**0.854**	0.8901	0.8904	0.9107	0.9687	0.8847
1007	1.1424	0.932	**0.8877**	1.0526	1.0807	1.2447	1.0057
1008	0.933	0.8813	0.9056	**0.8159**	0.9366	0.9958	0.9272
1009	1.0259	0.918	0.9044	0.8371	**0.8156**	0.8901	0.9275
10010	0.9697	0.9347	0.9035	0.8672	0.8963	**0.843**	0.9526
10011	1.0229	0.8924	0.8962	0.9041	0.9504	0.9508	**0.8564**

Table 3: Log scores for recognition of session 1 test data using combined session 1 and 2 model, using triphone durations.

3.2. Speaker recognition and accuracy

In order to quantify the value of temporal information in combination with acoustic information, we first tested a

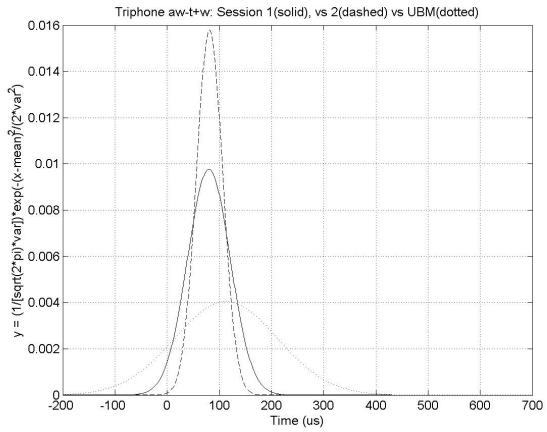

Figure 3: Speaker 10025 intersession variability for triphone aw-t+w in two different sessions, compared with the inter-speaker variability.

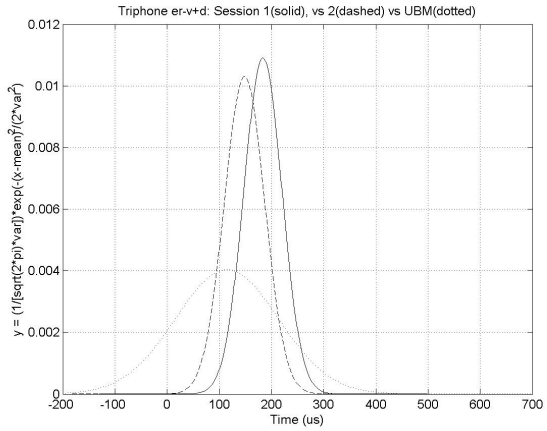

Figure 4: Speaker 1001 intersession variability for triphone er-v+d in two different sessions, compared with the inter-speaker variability.

system using only the forced alignment results (frame log likelihood probabilities). A threshold value was calculated for every speaker over the two sessions' training data, and a threshold value that was to be added to the speaker threshold value was computed in order minimize the number of errors made. The best possible threshold value was found to be 0.02 and resulted in 10 errors (7 FR and 3 FA). Of all the speakers in the pool, 16 were used for this test, giving a total of $2 \times (16 \times 16) = 512$ attempts at accessing the system. This gives an error percentage of $\frac{10}{512} = 2\%$.

Temporal information was then added to this system. The system subsequently achieved a best performance over two sessions of verification of 4 errors (3 FR and 1 FA). This gives an error percentage of $\frac{4}{512} = 0.78\%$. Hence, the error rate is reduced by a factor of 2.5 by the addition of temporal information.

	1001	1006	1007	1008	1009	10010	10011
1001	0.9348	0.9106	0.9157	0.979	0.9778	0.9266	0.946
1006	1.207	**0.8382**	0.979	1.2594	1.3615	1.3249	1.027
1007	1.1853	0.8714	**0.8538**	0.9361	0.9796	1.1057	0.911
1008	0.9146	0.8645	0.8864	**0.8024**	0.8656	0.9034	0.8838
1009	**0.8806**	0.8695	0.8984	0.8576	**0.8165**	0.9639	**0.8566**
10010	1.0369	0.954	0.9851	0.9926	1.0482	**0.842**	0.9807
10011	1.2568	0.8904	1.0009	1.0495	1.171	1.2449	0.8809

Table 4: Log scores for recognition of session 2 test data using combined session 1 and 2 model, using triphone durations.

3.3. Ability to detect recordings

Recordings replayed over a channel will introduce noise into the signal [5] and this will introduce variability into the timing of even identical utterances. We therefore need to compare the variability introduced by this process with the true intra-speaker variability. Hence, the duration of words in one verification session for a single speaker was investigated. As can be seen in figure 5, there is significant variation in the durations of the spoken words within a single session from one speaker.

For comparison, three sets of "noise-affected recordings" were simulated by randomly adding white Gaussian noise to every utterance; the noise level was chosen to give a signal-to-noise ratio (SNR) of 30 dB, in order to simulate realistic channel effects. Both the durations of whole words and phonemes were investigated (figures 6 and 7).

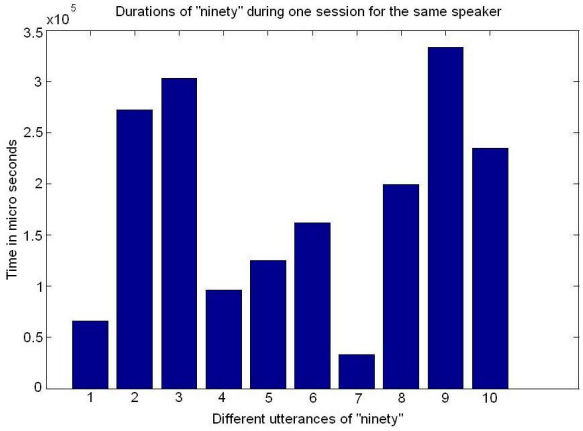

Figure 5: Different time durations for utterances of "ninety" by the same speaker in one verification session.

Figure 6 shows corresponding instances of utterances from 3 recordings of the same file grouped in threes. As can be seen, there is relatively large intra-speaker variation (even within a session) and little variability between

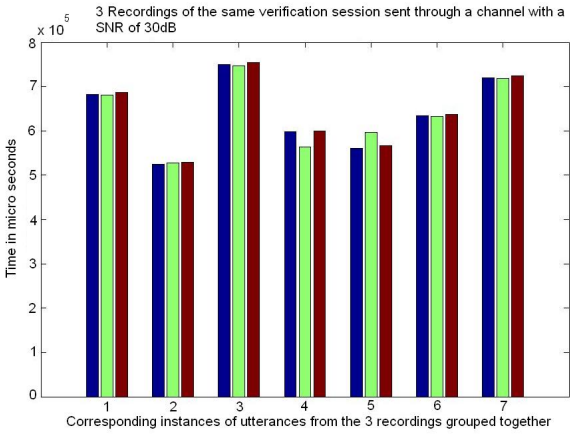

Figure 6: Corresponding durations of an utterance grouped together from three sets of the same verification session that was passed through a channel with a SNR of 30dB.

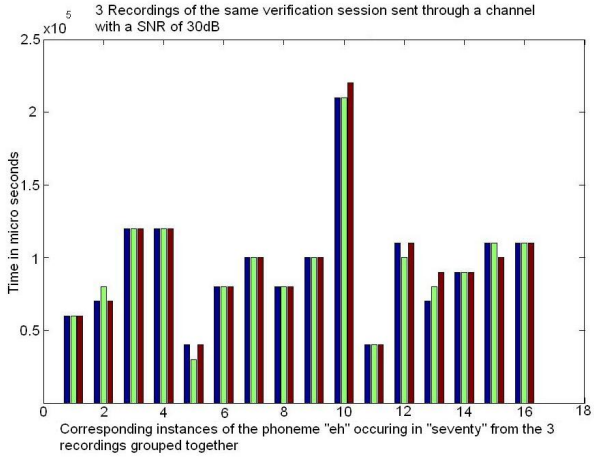

Figure 7: Corresponding durations of the phoneme ”eh” in s-e-venty grouped together from three sets of the same verification session that was passed through a channel with a SNR of 30dB.

corresponding words when noise is added.

The results suggest that the hypothesis regarding substantial intra-speaker variability and less variability when recordings were played is in fact true. Temporal information can thus be used effectively in detecting a recording.

4. Conclusion

Speaker verification is increasingly popular as a technique for implementing secure access control. It is generally accepted by the public as a non-intrusive biometric and is thus commercially attractive. Successful verification systems have been built around acoustic likelihoods [6], and our results suggest that such systems can be improved further by using temporal information.

To be effective against fraudulent attacks, these systems have to be robust against recordings being played to gain unauthorized access. Thus far, randomized phrase prompting has been the most popular approach. Our results suggest that one can protect against recordings by setting a threshold on the allowable similarity in identical triphone times, since the variability across utterances of the timings of words or phonemes produced by a live speaker is significantly larger than those induced by noise present in a recording.

These investigations should be extended in a number of directions. On the one hand, it is necessary to assess the magnitude of improvements attainable with temporal information on larger test sets, and in different verification or identification paradigms. It will also be very interesting to see the rejection rates achievable with realistic recordings, based on the measurements we have obtained. These topics are currently being investigated.

5. Acknowledgements

We are grateful to Dr. Marelie Davel for several useful insights.

6. References

[1] J. P. Campbell, “Speaker recognition: A tutorial,” *Proc. IEEE*, vol. 85, pp. 1437–1462, September 1997.

[2] “Otg white paper: Speaker verification in today’s marketplace,” 17 May 2005, `http://www.findbiometrics.com/Pages/voice%20articles/voice%20whitepaper%.pdf`.

[3] A. Higgins, L. Bahler, and J. Porter, “Speaker verification using randomized phrase prompting,” *Digital Signal Processing*, vol. 1, pp. 89–106, 1991.

[4] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland, “The HTK book. revised for HTK version 3.3,” September 2005, `http://htk.eng.cam.ac.uk/`.

[5] D.A. Reynolds, “Large population speaker identification using clean and telephone speech,” *IEEE Signal Processing Letters*, vol. 2, pp. 46–48, March 1995.

[6] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.