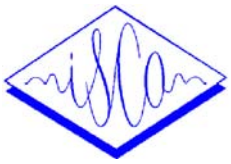

4th International Conference on Spoken Language Processing (ICSLP 96)
Philadelphia, PA, USA
October 3-6, 1996

GENERAL PHRASE SPEAKER VERIFICATION USING SUB-WORD BACKGROUND MODELS AND LIKELIHOOD-RATIO SCORING

S. Parthasarathy *A. E. Rosenberg*

AT&T Research, Murray Hill, NJ 07974

ABSTRACT

We present a design and study the performance of a text-dependent speaker verification system using general phrase passwords. The text of the password utterance and its phone transcription are assumed to be available. The problems that are addressed include the appropriate choice of units for building target speaker models and the choice of background models for likelihood-ratio scoring.

1. INTRODUCTION

The performance of speaker verification systems using digit-string passwords, evaluated on realistic telephone network data, has improved significantly in the last few years [1]. It is believed that, for some applications, passwords consisting of English phrases or even proper names is preferable to using digit-string passwords. A straight-forward extension of the digit-string system to handle phrases is to transcribe the speech utterance as a phone-string instead of a digit-string, and build target phone models instead of digit models. It is assumed that the phone transcription for the utterance is available. The transcription can be obtained either by looking-up the orthographic transcription in a dictionary or by submitting it to the front-end of a text-to-speech system. A phone-based speaker verification system is described in the next section.

Solutions to various problems that arise, essentially due to the fact that phones tend to be of shorter duration than digits and are more difficult to segment consistently, are presented in the rest of the paper. Target models for longer units, words and phrases, are studied in order to alleviate the segmentation problem. An energy based robust scoring method is also presented which makes the scoring less sensitive to segmentation errors. Another problem that has been studied by a number of researchers is the construction of background models [1, 2, 3]. A preliminary investigation to understand the importance of representing temporal details in the background model is presented.

2. PHONE-BASED SPEAKER VERIFICATION

The block diagram of a phone-based verification system is shown in Fig. 1. The user makes an identity (ID) claim by some means. Given an ID, the system expects a phrase associated with that ID at enrollment time. A speaker independent (SI) phone recognizer segments the input utterance into a sequence of phones that represents the phrase. The recognizer could reject the utterance if a significantly different phrase is uttered. The ID, input speech or feature vectors, and the sequence of phones with the corresponding end-points, are transmitted to the verifier. The feature vector is composed of 12 cepstrum and 12 delta cepstrum coefficients. The cepstrum is derived from a 10^{th} order LPC analysis over a 30 ms window. The feature vectors are updated at 10 ms intervals. The mean of the cepstral coefficients is estimated on an utterance-by-utterance basis using only the speech portions of the utterance. The verifier uses mean removed cepstrum.

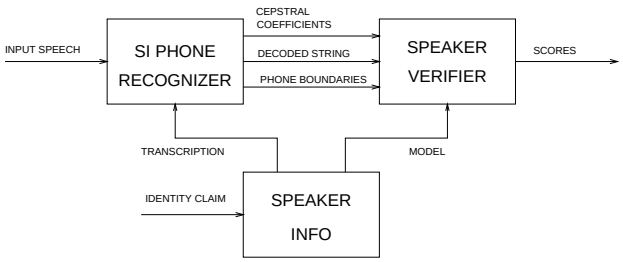

Figure 1: A phone-based speaker verification system.

At enrollment, a target speaker HMM is trained for each phone that appears in the transcription of the phrase chosen by that speaker. The verifier also has access to another set of phone models, called speaker background models, which can be generated in many ways [1, 2, 3]. Since the target speaker models are trained on a small amount of data from a single session, acoustic differences between training and test data tend to produce poor segmentation at the phone level. In this system, the phone end-points are constrained to

be those provided by the SI recognizer for computing both the target speaker score as well as the background score. The SI models tend to produce more consistent phone segmentation than the target speaker model. The state segmentations are determined by Viterbi decoding using the target speaker model for obtaining target scores and the background model to obtain the background scores. The verifier scores the input utterance using the target model and the background model. The likelihood test statistic is given by

$$L(\mathbf{O}, \Lambda_t) = \frac{1}{N_s} \sum_{i \in non-silence\ phones} p(\mathbf{O_i}|\Lambda_t)$$

where $p(\mathbf{O_i}|\Lambda_t)$ is the log-likelihood of the observation sequence O_i given the model Λ_t, $\mathbf{O}$ is the observation sequence over the whole phrase, $\mathbf{O_i}$ is the observation sequence that lies in the i^{th} phone, Λ_t is the set of target speaker phone models, and N_s is the total number of frames labelled as speech. The likelihood-ratio test statistic is defined as

$$L_r(\mathbf{O}; \Lambda_t, \Lambda_b) = L(\mathbf{O}, \Lambda_t) - L(\mathbf{O}, \Lambda_b)$$

where Λ_b is the speaker background model. In this paper, two sets of results are reported, one using $L(\mathbf{O}; \Lambda_t)$ and the other using $L_r(\mathbf{O}; \Lambda_t, \Lambda_b)$.

Adapting target models by updating them using the current accepted utterance has been found to be important for providing good verification performance [1].

3. DATABASE DESCRIPTION AND EXPERIMENTAL SETUP

This study was conducted using a phrase database collected over long distance telephone lines. The data used in this study are sentence length phrases that are common to every speaker. The results presented here are for the phrase "I pledge allegiance to the flag". Each speaker made 26 calls, each call being considered a session. Callers were instructed to use a variety of phones, locations, etc. so that a range of conditions are sampled over the different sessions.

The evaluation is done on a set of 100 speakers, 51 male and 49 female. Performance is measured by computing equal-error rates using *true* and *imposter* testing data. Two error rates are computed: one using $L(\mathbf{O}; \Lambda_t)$ as the test statistic and the other using $L_r(\mathbf{O}; \Lambda_t, \Lambda_b)$ as the test statistic. The quoted error rates are an average of the individual equal error rates.

Five repetitions of the phrase recorded in an enrollment session is used for training. An HMM is trained for each phone that appears in the transcription of the speaker's phrase. All the phone HMMs are 3-state, left-to-right, with a mixture Gaussian probability density function (*pdf*) associated with each state. Each state *pdf* had nominally 4 mixture components. As has been observed previously ([1]), the variance estimates are unreliable due to limited amount of training data. Therefore, a global variance estimate is used that is common to all the mixture components.

For evaluation, the data is divided into *true* and *imposter* tokens. The *true* tokens consist of 2 repetitions of the phrase in each of 25 test sessions, for a total of 50 tokens. The *imposter* data consists of 4 tokens from every speaker of the same gender as the true speaker, which is roughly 200 tokens.

The choice of a speaker background model can significantly affect the verification performance. In this paper, a single set of phone models is used as the background model for all the target speakers. The background model is trained on a telephone speech database obtained from an independent population of speakers and using speech material that is different from the phrase database described above.

Experiments are carried out using unadapted models as well as models adapted using four target speaker verification utterances sequentially. Imposter utterances are tested against fully adapted target models.

The first row of Table 1 shows the four equal error rates for the system described above. The first two colums are unadapted results and the other two are adapted results. A few points are worth mentioning.

- The L_r statistic reduces the error rate by almost 50% over the error rate using the unnormalized statistic L, with adapted as well as unadapted target speaker models. This demonstrates that the SI phone models are quite effective as speaker background models.

- Sequential adaptation on 4 test utterances reduces the error rate by 14% using L and 20% using L_r.

- The best adapted error rate is approximately 3 times the previously reported error rate for digit-string password systems [1]. Since these are different databases, it is difficult to compare the results, but such a large disparity needs to be explained. The following sections address some of these issues.

4. HYBRID WORD-PHRASE-PHONE REPRESENTATIONS

In the system described in Section 2, the phrase is transcribed as a sequence of phones. This was done in order

to make sure that the system was general enough to handle any English phrase that the user may choose. However, phones are shorter in duration than words and the phone-level segmentation is generally less consistent than word-level segmentations. This inconsistency could result in degraded performance, especially when the models are adapted. For example, in the case of digit string systems, adaptation typically reduces the error rate by much more than the 20% observed in the baseline results [1].

One way of realizing some of the benefits of word-level representations without sacrificing generality is to build target speaker models for each word in the user's phrase, but continue to represent the phrase as a sequence of phones for scoring with background models.

Each word in the phrase is represented by a left-to-right HMM with roughly 1.5 times the number of states as there are phones in the word. The results obtained using this hybrid word-phone system are shown in row two of Table 1. Some observations:

- The unadapted results are only a modest modest improvement over the baseline indicating that there is no significant gain due to reduced number of states or the use of larger units.
- The adapted results are significantly better. The error rates using L_r with adaptation (the last column in Table 1) is usually of most interest since that is the mode in which practical systems are likely to operate. This error rate is a 18% reduction over the baseline suggesting that the inconsistency of the phone end-points prevented effective adaptation and degraded the performance of the system.

In the word-phone system above, the words were taken to be lexical items in the orthography for the phrase. Such a definition was just a convenience rather than a necessity. Short words, such as "I", "to", and "the", in the example phrase "I pledge allegiance to the flag" may be just as difficult to segment as phones. Target speaker models could be built for groups of words by combining short or function words with their neighbours. This may improve the performance beyond what was obtained by the word-phrase system described earlier.

In the limit, it is also possible to represent the entire phrase by a single HMM. The disadvantage is that long pauses within a phrase become harder to model. However, for phrases of interest which tend to be approximately two seconds in duration or less, there is unlikely to be long pauses.

An evaluation was done by building a target model consisting of a single HMM for the user's phrase and using the SI phone models for background scoring. The results are shown on the third row of Table 1 and demonstrate the following:

- The most significant observation is that the adapted, L_r result is 1.9% which is a 29% reduction over the baseline.
- The improvement using unadapted models and L_r is only about 15% over the baseline.
- It is also interesting to note that the improvement over the baseline using the L test statistic is very modest in the case of adapted models and none at all without adaptation. The general conclusion is that the segmentation effects are significant but are noticeable only when the target models are trained on a reasonable amount of data (adapted results) and when the test statistic is reasonable (L_r).

5. ENERGY BASED ROBUST SCORING

An analysis of the phone segmentation indicated, that in some cases, short silence segments were included as part of phone segment. This typically happened at word or phrase boundaries. One way of reducing the variance of the likelihood estimates due to the inclusion of these silence frames is to not include them in the scoring. A simple algorithm for accomplishing this using energy-based prunning is described below.

The first step is to establish an estimate of the background energy level on an individual utterance basis. This is done by computing an normalized energy histogram over the utterance and selecting the background level to be the energy value at which the cumulative histogram exceeds a predetermined level.

Consider the system where the target model is a single HMM for the whole phrase. Let the per-frame log-likelihood of the observation sequence on the best Viterbi path be given by $l_0^t, l_1^t, ..., l_{N-1}^t$ where N is the number of frames labelled as the phrase, and l_i^t is the likelihood of the observation $\mathbf{o}_i$ scored by the target model. The statistic L_r is given by

$$L_r = \frac{1}{N_e} \sum_{i \in e_i > T} (l_i^t - l_i^b)$$

where N_e is the number of frames whose energy exceeds the threshold T, and e_i is the energy of the i_{th} frame.

The results using the robust scoring procedure described above is shown in Table 2. The first row is the same as the adapted results from row 3 of Table 1. The second row shows the improvement using robust

Target model	equal-error rates (%)			
	without adaptation		with adaptation	
	using L	using L_r	using L	using L_r
Phone	6.2	3.4	5.3	2.7
Word	5.9	3.2	4.8	2.2
Phrase	6.3	2.9	4.9	1.9

Table 1: Performance of the phrase verification system in terms of the average individual equal error rate in %. The background model in all cases is a SI phone model trained on an independent database.

Condition	equal-error rates (%)	
	using L	using L_r
Phrase target, phone background	4.9	1.9
Phrase target, phone background, robust	4.3	1.8
Phrase target, 1 state background, robust	4.3	4.0

Table 2: Performance of the system using phrase target model and with adaptation under various conditions.

scoring. The improvement overall is modest. Analysis of error rates for individual speakers shows that robust scoring never degrades results significantly but improves significantly for a small number of speakers. Since it costs so little to do the robust scoring, even a small improvement justifies its use.

6. PRELIMINARY EXPERIMENTS WITH SIMPLE BACKGROUND MODELS

In all the experiments so far, SI phone models have been used to obtain the background score. It is not clear if such a detailed representation is necessary to obtain an effective background model. There are many ways of building simpler background models. An extreme control condition is to simply collapse the background model to a single state with a mixture *pdf* associated with it. This scheme would eliminate the necessity for obtaining phone transcriptions altogether. A simple experiment was conducted where a single state 128 mixture model was trained on the same database that the SI phone models were trained on. The results are shown on row 3 of Table 2. The error rate using L is unchanged from row 1 since the background model in not used in computing L. However, the results using L_r are much worse than the corresponding result in row 2, demonstrating that it is important to incorporate temporal and linguistic constraints in the background model. This experiment is preliminary and further studies are needed to fully understand the requirements on the background model.

7. CONCLUSION

A speaker verification system using general English phrase passwords has been developed. The baseline system was simply an extension of the digit-string password system that had been previously developed. It was shown that using word or phrase target models provides much better performance than using phone target models. A technique for robust scoring using energy based frame selection was shown to be useful. Use of phrase target models and robust scoring provided a 33% reduction in the equal error rate over the baseline. Some preliminary evidence was presented to show that it is useful to incorporate some temporal structure in the background models.

8. REFERENCES

[1] A. E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password verification," *to appear in ICASSP-96*, 1996.

[2] M.J. Carey, E.S. Parris, and J.S. Bridle, "A Speaker Verification System Using Alpha-Nets," *Proc. Int. Conference on Acoustics, Speech, and Signal Processing*, pp. 397-400, Toronto, 1991.

[3] T. Matsui and S. Furui, "Similarity Normalization Method for Speaker Verification Based on a Posteriori Probability," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification,* pp. 59-62, 1994.