SPEAKER BACKGROUND MODELS FOR CONNECTED DIGIT PASSWORD SPEAKER VERIFICATION

Aaron E. Rosenberg *S. Parthasarathy*

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974

ABSTRACT

Likelihood ratio or cohort normalized scoring has been shown to be effective for improving the performance of speaker verification systems. An important problem in this connection is the establishment of principles for constructing speaker background or cohort models which provide the most effective normalized scores. Several kinds of speaker background models are studied in this paper. These include individual speaker models, models constructed from the pooled utterances of different numbers of speakers, models selected on the basis of similarity with customer models, models constructed from random selections of speakers, and models constructed from databases recorded under different conditions than the customer models. The results of experiments show that pooled models based on similarity to the reference speaker perform better than individual cohort models from the same similar set of speakers. Pooled background models from a small number of speakers based on similarity perform about the best, but not significantly better than a random selection of 40 or more gender balanced speakers with training conditions matched to the reference speakers.

1. INTRODUCTION

The use of likelihood ratio or cohort normalized scoring has improved the performance of many speaker verification systems[1, 2, 3]. In likelihood ratio scoring an utterance is scored not only against a specified customer reference model but also against a so-called set of cohort models or a speaker background model. The verification score is calculated as the ratio of the customer reference and speaker background likelihoods. In terms of log likelihoods, the normalized verification score is calculated as the difference of log likelihoods, as follows:

$$\log L(\mathbf{O}) = \log p(\mathbf{O}|S = I) - \log p(\mathbf{O}|S \neq I) \qquad (1)$$

where $\mathbf{O}$ represents the observed measurements for an utterance, $p(\mathbf{O}|S = I)$ is the likelihood of the measurements with respect to the model for the customer I and $p(\mathbf{O}|S \neq I)$ is the likelihood of the measurements for speakers other than I. The likelihood ratio score tends to be more stable and less variable than the unnormalized reference model score leading to improved performance. The question considered here is how to construct models which provide the normalizing or speaker background likelihood. There have been many suggestions for constructing speaker background models[1, 2, 3, 4] but general principles for construction leading to optimum performance are lacking.

In this report different kinds of background model constructions are studied. The term cohort model is reserved for a member of a set of *individual* reference models for which a normalizing score is obtained. That is the normalizing score is obtained as

$$p(\mathbf{O}|S \neq I) = \frac{1}{K}\sum_{k=1}^{K} p(\mathbf{O}|c_k(I)) \qquad (2)$$

where $c_k(I)$, $k = 1, 2, \ldots, K$ is an individual speaker model whose selection depends on the customer model I. The term background speaker model is used to refer to a single model, constructed by pooling training utterances from more than one speaker, whose selection, in general, might depend on I

$$p(\mathbf{O}|S \neq I) = p(\mathbf{O}|\lambda(I)) \qquad (3)$$

The various speaker background models studied in this paper include individual speaker models, models pooled over different numbers of speakers, models selected on the basis of similarity with customer models, random selections of speaker models, and models from databases recorded under different conditions and applications than the customer models.

2. HMM-BASED SPEAKER VERIFICATION

Experiments are carried out using a word-based, continuous density Gaussian mixture, hidden Markov model (HMM) speaker verification system. The system is operated in a text-dependent mode with customers assigned fixed digit-string passwords. Customers are represented by a set of HMM digit models comprising the digits in their password utterances. Each digit model contains a left-to-right sequence of 6 Markov states. Each state contains a maximum of 4 mixture components. Speaker background models also contain 6 states per digit. However, pooled models contain 12 mixture components per state while individual models in a cohort contain 4 mixture components per state. Model parameters are estimated by means of a segmental K-means training procedure[5]. A set of fixed variance parameters is used to replace estimated model variances for both customer and speaker background models. Speaker independent digit HMM's are used to provide segmentations of utterances into digit and silence or background segments.

0-7803-3192-3/96 $5.00©1996 IEEE

These segmentations are used as initial segmentations for training utterances and to specify scoring segments for verification utterances.

Verification scores are obtained by comparing the segmented verification utterances with customer and speaker background models. A log likelihood scores is obtained for each digit segment using Viterbi decoding to find an optimal state segmentation for the associated digit HMM. An average log likelihood score is obtained over all digit segments normalized by the total length of the utterance. The normalized verification score is the difference of the reference model and background model log likelihood scores.

Adapting customer models by updating them using current customer verification utterances has been found to be important for providing good verification performance. In this study, experiments are carried out on both unadapted customer models and adapted models updated using 4 customer verification utterances.

3. EXPERIMENTAL DATABASE

The database consists of 14-digit verification password utterances collected over the long distance network. The database contains two sets of speakers, an evaluation set and an auxiliary set. The evaluation set consists of 49 adult speakers designated customers, 25 female and 24 male plus a set of imposters. Each speaker designated as a customer provided 3 enrollment utterances of the password recorded in a single call. Subsequently, each speaker provided a series of verification utterances, 1 or 2 per call. The speakers were encouraged to use a variety of telephones and conditions, but to avoid speakerphones. The total number of verification utterances per customer ranges from approximately 35 to 150. The average is 65. To provide the imposter set, for each customer, a set of 30 or 35 speakers of the same gender each recorded 2 utterances of the customer's password in a single session. (Occasionally only 1 utterance was recorded in a session.) The average number of imposter utterances per customer is 65.

An auxiliary set of speakers, 36 female and 40 male, distinct from the evaluation set, each provided 3 enrollment utterances. These utterances are used to construct speaker background models.

The utterances are digitized with a 3200 Hz lowpass antialiasing filter. The digitized utterances were input to a 300 Hz highpass filter, preemphasized with a first order difference digital network, and converted to 10th order linear predictive coding (LPC) coefficients every 10 ms over 30 ms windows throughout each utterance. The LPC coefficients are converted to 12th order cepstral coefficients and augmented by 12th order delta cepstral coefficients calculated over a 5-frame window of cepstral coefficients. In addition, blind channel equalization is carried out by calculating the average of each cepstral coefficient over the speech portions of each utterance and subtracting it from the instantaneous cepstral coefficients for each analysis frame.

4. SPEAKER BACKGROUND MODELS

The following experimental conditions are studied for the realization of models which provide the normalizing term in the log likelihood ratio verification score formulation in Eq. 1.

- model configuration: individual speaker models or pooled speaker model
- selection criterion: based on similarity to reference model or random selection
- model population size
- model population source

4.1. Individual speaker background models based on similarity

In our previous study[2] the normalizing term is associated with a set of individual speaker models selected on the basis of a similarity measure between each model and the true speaker model. The speaker set is referred to as the cohort assigned to the customer. The similarity measure is based on the log likelihood of mean vectors of the first speaker model for the state model parameters of the second speaker model. The cohort models are obtained by first measuring the similarity of each speaker's digit model in the auxiliary database to the customer's digit model for each digit in the customer's account number password to produce a list of auxiliary speaker models sorted by similarity for each digit. The best matching models are merged to form a complete set of digit models for the true speaker's password. Although the merged set is treated as if it were a single speaker model set, in general, the models in each set originate from different speakers. Next, the second best matching models are merged into a single set of models and the process continues up to the nth best matching models where n is set to 5 for these experiments.

In general, the speakers that make up each set of cohort models are distinct from one true speaker to another but, almost universally, they are the same gender as the true speaker.

Let $c_k(I)$, $k = 1, 2, \ldots, 5$ be the the best, second best, up to 5-th best merged cohort model sets for speaker I. The normalizing score is the average likelihood score over the cohort models as shown in Eq. 2.

4.2. Pooled speaker background models based on similarity

To compare the use of individual speaker models and pooled speaker models for the normalizing term, a set of speaker dependent, pooled background speaker models are constructed with the same speaker composition as the individual cohort models described in the previous section. In other words, the training utterances corresponding to the best matching cohort sets obtained for a customer are pooled to create a single pooled background speaker model tailored for that customer. Recall that the pooled models contain 12 mixture components per state while the individual speaker models contain, nominally, 4.

4.3. Pooled speaker background models with random speaker compositions

Finally, to evaluate the effects of size, composition, and database source on pooled speaker background models, the following pooled speaker background models are constructed.

Varying the speaker population size of pooled speaker models. Pooled speaker background models are constructed from the training utterances of the speakers in the auxiliary database. The population sizes are 5, 20, and 76. As population size decreases, performance becomes more sensitive to the speaker composition in the background model. That is, for small size populations the performance for a given true speaker may vary significantly from one pooled speaker background model to another depending on the composition of the model. The sensitivity is a function of both gender and individual same-gender differences. To smooth these effects in the evaluation of overall performance two 20-speaker background models are constructed, each with a distinct, gender balanced, composition of speakers. For size 5 models, single gender models are constructed. Male background models are used for male true speakers and female background models are used for female true speakers. Four distinct size 5 models are constructed for each gender. Performance for size 5 and size 20 background models are obtained by averaging the results obtained with the four size 5 models and the two size 20 models.

Pooled speaker models from other databases. Two additional pooled background models are constructed from different telephone network databases than the one used for evaluation. The first is a 39-speaker database of speakers using 9-digit passwords. The model is constructed from the enrollment utterances which contain 3 tokens of each speaker's password. The database is approximately gender balanced. The second database is constructed from an approximately 100-speaker database of 10-digit utterances. Each speaker recorded one utterance. The recording setup for this database is different than the one used for the evaluation database and the 39-speaker database.

5. RESULTS

Speaker verification experiments are carried out for each of the background model specifications described in the previous section. In each experiment, for each true speaker, the set of true speaker test utterances are scored against the true speaker model and specified background model to obtain a set of true speaker scores. Also, the set of imposter test utterances are scored against the true speaker model and specified speaker model to obtain a set of false speaker scores for the speaker.

A priori thresholds are not assigned. Instead, individual and pooled equal-error rates are calculated. Equal-error is calculated by sorting true and false speaker test scores and finding the score value such that the fraction of true scores less than that value is equal to the fraction of false scores greater than that value. This fraction is known as the equal-error rate since, if the decision threshold were actually set at that value, the experimental outcome would be that the false reject rate would equal the false accept rate. Two equal-error rate calculations are carried out. The first is an individual speaker equal-error rate. The performance figure is obtained by averaging all the evaluation speaker individual equal-error rates. The second is a pooled equal-error rate. This is obtained by pooling the true speaker scores and the false speaker scores for all evaluation speakers and calculating an equal-error based on a single, speaker independent threshold. Performance measured on the basis of individual equal-error rates is always better than pooled equal-error rate performance. Since score ranges generally vary from speaker to speaker, individual thresholds can provide better performance in speaker verification than speaker independent thresholds.

Results are reported for both unadapted and adapted true speaker models. Supervised adaptation is carried out in the following way. The true test utterances are compared and scored in the order in which they were recorded. Every other true test utterance is used to update the true speaker models until four updates are completed. After the model updating is completed, the false test utterances are scored. Model updating is carried out by using test utterance data to update mixture component mean vectors and weights in a technique similar to that described in Rosenberg and Soong[6].

The results are summarized in Table 1. First, observe that the performance rankings across the four different performance categories are largely in agreement. For example, the best performance for 3 out of the 4 categories is "5 best pooled" and the worst performance is "5 random pooled" for 4 out of the 4 categories. The range of performance for average individual equal-error rates is from 2.2% to 3.4% for unadapted models and 1.0% to 1.4% for adapted models. For pooled equal-error rates the ranges are 5.0% to 6.3% for unadapted models and 2.1% to 3.2% for adapted models. Roughly speaking, the range is 40% to 50% from worst to best performance for all categories.

The first three rows in the table show the results for 5-speaker background models. The training utterances here are all drawn from the auxiliary speaker set. The best results are obtained by pooling the training utterances for the 5 most similar digit models for each true speaker (row 2); the worst results are obtained for random selections of 5 speakers (row 3). Using the 5 most similar individual speaker digit models (row 1), the original cohort concept, the performance is also relatively poor.

Rows 3 through 5 show results for pooled models of different size constructed from random selections of speakers from the auxiliary set. It can be seen that performance improves monotonically with speaker set size.

Finally, rows 5 through 7 show results for pooled models from different database populations, all drawn from large size speaker sets. The differences in performance are small. Somewhat surprisingly, the performance for the 39-speaker population (row 6) is as good or better than the 76-speaker population from the same database (row 5). The 500-speaker population results (row 7) are slightly worse, perhaps because the utterances were recorded in a different setup.

The Wilcoxon Matched-Pairs Signed-Ranks Test of statistical significance [7] is used to test whether or not performance for a particular background model condition is significantly different than another. The test is applied to equal-error rate differences for each individual between two conditions. Looking at the results with adaptation as an example, these tests indicate no statistically significant differences among the the four best performances (rows 2, 5,

row description	equal-error rates (%)			
	without adaptation		with adaptation	
	avg. indiv.	pooled	avg. indiv.	pooled
1 5 best individual	2.96	5.64	1.34	2.66
2 5 best pooled	2.23	5.08	1.02	2.10
3 5 random pooled	3.36	6.30	1.44	3.23
4 20 random pooled	3.21	6.00	1.36	2.97
5 76 random pooled	2.75	5.40	1.04	2.26
6 39 pooled database	2.23	5.04	1.09	2.20
7 500 pooled database	2.57	5.48	1.11	2.47

Table 1. Equal-error rate performance, without and with adaptation, for various speaker background model conditions.

6, and 7) and among the three worst performances (rows 1, 3, and 4). On the other hand, between members of each group, the differences are weakly statistically significant (at the 5% level) or better.

6. DISCUSSION AND CONCLUSION

From these experiments, it can be concluded that pooled background models based on model similarity perform better than individual background models when constructed with the same speakers. For pooled background models, a speaker dependent selection of a small number of speakers whose utterances are similar to the reference speaker's model performs about the best, but not significantly better than a random selection of 40 or more gender balanced speakers selected from a database of speakers with matched recording conditions. It should be realized that for the pooled models based on similarity, since the selection is speaker dependent, a separate background model must be constructed for each reference speaker. Thus, any advantage in performance might be outweighed by this additional complexity.

General theoretical principles for constructing good background models are not known. But certain properties have been observed which are consistent with effectiveness. First, background scores should be strongly correlated with reference scores for true and imposter scores, but especially for true scores. Second, there should be a strong tendency for true reference model scores to be relatively higher than true background model scores and vice versa for imposter scores.

REFERENCES

[1] A. Higgins, L. Bahler, and J. Porter, "Speaker Verification Using Randomized Phrase Prompting," *Digital Signal Processing*, vol. 1, pp. 89-106, 1991.

[2] A.E. Rosenberg, J. DeLong, C-H. Lee, B-H. Juang, and F.K. Soong, "The Use of Cohort Normalized Scores for Speaker Verification," *Proc. ICSLP 92*, vol. 2, pp. 599-602, International Conference on Spoken Language Processing, Banff, 1992.

[3] M.J. Carey, E.S. Parris, and J.S. Bridle, "A Speaker Verification System Using Alpha-Nets," *Proc. ICASSP 91*, pp. 397-400, International Conference on Acoustics, Speech, and Signal Processing, Toronto, 1991.

[4] 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.

[5] L.R. Rabiner, J.G. Wilpon, and B-H. Juang, "A Segmental K-Means Training Procedure for Connected Word Recognition," *AT&T Tech. J.*, vol. 65, pp. 21-31, 1986.

[6] A.E. Rosenberg and F.K. Soong, "Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes," *Computer Speech and Language*, vol. 2, pp. 143-157, 1987.

[7] S. Siegel, *Nonparametric Statistics for the Behavioral Sciences*, pp. 75-83, McGraw-Hill, 1956.