BACKGROUND MODEL DESIGN FOR FLEXIBLE AND PORTABLE SPEAKER VERIFICATION SYSTEMS

Olivier Siohan *Chin-Hui Lee* *Arun C. Surendran* *Qi Li*

Bell Labs, Lucent Technologies
600 Mountain Ave, Murray Hill NJ 07974.
email: {siohan,chl,acs,qli}@research.bell-labs.com

ABSTRACT

Most state-of-the-art speaker verification systems need a user model built from samples of the customer speech, and a speaker independent (SI) background model with high acoustic resolution. These systems rely heavily on the availability of speaker independent databases along with *a priori* knowledge about acoustic rules of the utterance, and depend on the consistency of acoustic conditions under which the SI models were trained. These constraints may be a burden in practical and portable devices such as palm-top computers or wireless handsets which place a premium on computation and memory, and where the user is free to choose any password utterance in any language, under any acoustic condition. In this paper, we present a novel and reliable approach to background model design when only the enrollment data is available. Preliminary results are provided to demonstrate the effectiveness of such systems.

1. INTRODUCTION

Speaker verification is the task of automatically determining whether the claimed identity of a speaker is correct, given some speech observations [1]. State-of-the-art fixed phrase speaker verification systems verify the identity of the speaker through a Neyman-Pearson test based on a normalized likelihood score of a spoken pass-phrase [2]. If λ_c is the customer model, given some acoustic observations X, the normalized score $s_{norm}(X, \lambda_c)$ is usually computed as the ratio of the likelihoods as follows:

$$s_{norm}(X, \lambda_c) = \frac{p(X; \lambda_c)}{p(X; \lambda_B)}, \quad (1)$$

where $p(X; \lambda)$ is the likelihood of the observations X given the model λ, and λ_B is a so-called background model. The customer model is usually a hidden Markov model (HMM) built from repeated utterances of a pass-phrase spoken by the customer during enrollment. This model is usually created either by concatenating phone-based customer HMMs or by directly estimating a whole-phrase HMM [3].The background model is usually an HMM that reduces the need for a speaker dependent threshold, and is built by concatenating speaker independent phone models of the customer password. In applications where it is desirable to give the customer the freedom to select his own password phrase, most traditional systems assume that the phonetic transcription of customer password is available, which in turn assumes the availability of pre-trained multi-lingual phone models, dictionaries and a set of letter-to-sound rules for the particular language. Because good phone end-points are necessary, a speaker independent phone recognizer might be used to derive the phone segmentation. The whole architecture of such speaker verification system can therefore be quite complicated (cf. Fig 1). Furthermore, a good set of SI background phone models often imply each model to have a large acoustic resolution, i.e., more mixture components per state, in order to have a good performance (e.g. [3]). This means a higher demand of computation and memory requirement which may not be desirable for applications running on hand-held devices such as personal digital assistants, palm-top computers or wireless phones. There is also an issue of robustness - the background SI phone models provided by the system may exhibit very different acoustic properties from the operating condition. For practical and portable applications, the previous requirements are not acceptable and are a burden on the customer and the system developer. The customer may want to select a password in any language, may choose to perform speaker verification with any microphone under any acoustic condition.

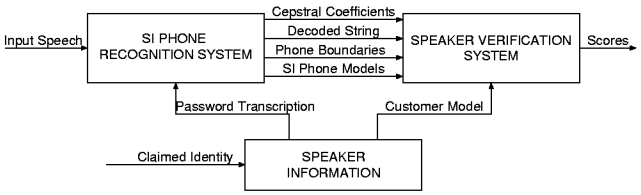

Figure 1: Architecture of a traditional phrase based speaker verification system (adapted from [3]).

Our goal in this work is to build a flexible, portable

speaker verification system which minimizes the constraints on the customer and simplifies the whole system architecture. We assume that the only speech material available is the set of enrollment utterances provided by the customer, that the customer is free to select a password phrase in any language, and that no speaker independent models are available. We will show that even though the system is built with no prior information, the performance is appreciably good, though it does not match that of a system using very high resolution SI phone models. But the focus indeed is on the flexibility and simplicity provided by our novel approach. The principle of such a system is explained in the next section.

2. PRINCIPLE

Our working assumption is that the customer will choose his/her own password phrase and will be asked to repeat this utterance several times for enrollment. No other speech data or model is available, nor orthographic or phonetic transcription of the password utterance.

The acoustic information in the customer password phrase is modeled using a whole-phrase HMM, λ_c, estimated from the enrollment utterances. One disadvantage of using a whole-phrase model is that pauses within the phrase can be hard to model, and might upset the decoding and computation of $p(X;\lambda_c)$. However, in most practical situations, password utterances are very short (less than 2 seconds) and long pauses are unlikely.

Typical state-of-the-art speaker verification systems build background models from speaker independent databases. Some studies advocate that the background model λ_B should be derived from speakers randomly selected from a speaker independent database [4]. Others suggest to select speakers that are "close" to the customer, and are therefore representative of the population near the claimed speaker (cohort speakers) [2, 5], which is expected to improve the selectivity of the system against voices similar to the customer. Such a scenario is impractical for portable applications since a speech database would have to be provided to the customer.

Since cohort modeling emphasizes the use of speakers having acoustic characteristics similar to the customer in order to train the background model, we propose to build the background model λ_B directly from the customer enrollment utterances. Of course, to end-up with a background model λ_B different from the customer model λ_c, it is necessary to perturb the model λ_B after or during training, or to perturb the enrollment data before estimating the model. This can be done in many ways, for example by adding noise to the enrollment utterances before estimating λ_B, or by perturbing the variance of the background model after training.

In this paper, we propose two different techniques to perturb λ_B. The first one is to use a background model λ_B having a small number of parameters compared to λ_c, and therefore a cruder acoustic resolution. The model λ_c is designed to provide a fine acoustic resolution of the password-phrase, while the model λ_B only provides a rough acoustic resolution. This can be done by using a small number of states for λ_B compared to λ_c. The architecture of the proposed system is therefore very simple, and is represented in Fig. 2. λ_B acts like an anti-customer model, and each customer has his/her own anti-model. The customer model λ_c consists of a large number of states, say 25, while the background model consists of a small number of states, say less than 5. With such an architecture, the customer can choose any password in any language, no speaker independent data is used, and some useful acoustic normalization is provided with the background model. Since both models λ_c and λ_B are trained from the same data, there is no acoustic mismatch related to the environment between these two models. That contrasts with systems using pre-trained background models which are rather sensitive to changes in the environment, and might need to impose some additional constraints on the user, like for example the use of a particular microphone.

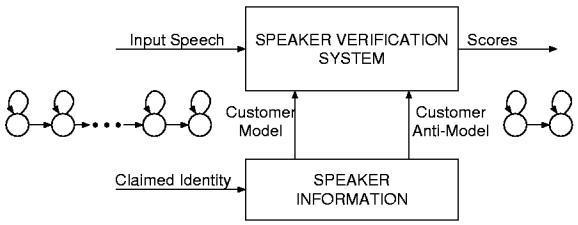

Figure 2: Proposed architecture of a phrase speaker verification system.

The second technique is to train a background model as previously described, and to reverse the state order after training, for example state 5 becoming state 1, state 4 becoming state 2, and state 3 being unchanged for a 5 states background HMM. By perturbing the temporal information, but still keeping a spectral information related to the customer model, we obtain a background model more "smear" that can still provide some acoustic normalization when computing the likelihood ratio.

3. EXPERIMENTS AND RESULTS

3.1. Database Description

The experimental evaluation is carried out using a database of spoken phrases recorded digitally over the telephone network. Volunteers and paid subjects recorded utterances from their home, office or other phones by dialing a toll-free number, and were encouraged to use a variety of phones, exclud-

ing speakerphones, so that a range of conditions are sampled over the different recoding sessions. The data used for the purpose of this evaluation consists of a phrase common to all speakers ("I pledge allegiance to the flag").

The database contains utterances spoken by 100 speakers, 51 males and 49 females. Each speaker provided 5 tokens of the common phrase in a single training session. Two tokens of the common phrase were also recorded in each of 25 testing sessions. Consequently, a total of 50 test utterance tokens are available from each speaker. For each speaker, 2 sessions (4 utterances) of the same sex speakers are also used as imposter test data, which is an average of 200 imposter utterances per speaker.

3.2. Front-end Processing

The signal is first passed through a 3200Hz low-pass anti-aliasing filter. A 300Hz high-pass filter is then applied to minimize the effect of processing in the telephone network. The resulting signal is pre-emphasized using a first order difference and 10th order linear predictive coding (LPC) coefficients are derived every 10ms over 30ms Hamming windowed segments. The 10 LPC coefficients are converted to 12th order cepstral coefficients (LPCC) and a feature vector of 24 components, consisting of 12 LPCC and their first derivatives is produced at each frame.

3.3. System Description

The speaker verification system is operated in a text-dependent mode. For evaluation purpose, all customers have the same password phrase, but the system architecture is such that each speaker could select his own password. Three models, λ_c, λ_B and λ_{sil} are built for each customer. The detailed model λ_c is a left-to-right HMM, consisting of 25 states with up to 4 Gaussian mixture components per state. The crude background model λ_B is a 5-states, left-to-right HMM with 4 Gaussian mixture components per state. A silence model λ_{sil}, consisting of 3 states and 4 Gaussian mixture components is also trained for each customer. All Gaussian probability density functions have a diagonal covariance matrix. Some experiments have been carried out using an even cruder background model with only 1-state and 4 mixtures, which is simply a Gaussian mixture model. All models are trained using the segmental $\mathcal{K}$-means algorithm [6], followed by 5 iterations of the EM algorithm, using the 5 training utterances. The covariance matrices of the detailed model λ_c are tied in order to get a more robust estimation.

For each test utterance, a Viterbi decoding is performed using the detailed model λ_c and the silence model λ_{sil} to find the optimal state segmentation and get a speech/silence segmentation. The speech segment is also decoded using the background model λ_B. Average log-likelihood scores $\log p(X|\lambda_c)$ and $\log p(X|\lambda_B)$ are obtained over the speech segment for the two models λ_c and λ_B.

EER	Raw Score	Normalized Score
Male	7.63	4.67
Female	9.38	6.59
Average	8.49	5.61

Table 1: Average individual equal error rates (EER). 25-states customer model, 1-state background model, 4 mixtures per states. Both customer and background model are training using $\mathcal{K}$-means followed by 5 EM iterations

Two verification scores are used to evaluate the performance of the system. The first one, $\tilde{s}_{raw} = \log p(X|\lambda_c)$, called "raw" score, involves only the log-likelihood derived from the model λ_c. The second one, $\tilde{s}_{norm} = \log p(X|\lambda_c) - \log p(X|\lambda_B)$ is the normalized score corresponding to the log-likelihood ratio. *A priori* thresholds are not assigned in our experiments and the system performance is evaluated from average individual equal-error rates. Equal-error rate is calculated by sorting customer and imposter verification scores and finding the score value such that the fraction of customer scores less than that value is equal to the fraction of imposter scores greater than that value. This fraction is the equal-error rate, meaning that if the decision threshold is set to that score value, the false rejection rate is equal to the false acceptance rate. The equal-error rate is derived individually for each speaker and averaged over male and female speakers to get an average individual equal error rate.

3.4. Results

In a first set of experiments, we used a 1-state background model λ_B with 4 Gaussian mixture per state. We recall that the customer model λ_c is a 25-states HMMs, using tied diagonal covariance matrices. The average equal error rate is given in Table 1 for male and female speakers, using the raw and normalized scores. The normalized score provides a significant improvement over the raw score, and illustrates that a background model can be built without any speaker independent data.

In a second set of experiments, we increased the number of states in the background model to check whether incorporating a temporal information in the background model would improve the performance. Results in table 2 are obtained using a 5-states background model. The raw scores are the same as in table 1 since the raw score does not involve the background model. A degradation of the EER is obtained, compared to the single state background model. That contrasts with systems using speaker independent background models where temporal information in the background model provide a significant improvement over a 1-

EER	Raw Score	Normalized Score
Male	7.63	5.62
Female	9.38	7.15
Average	8.49	6.37

Table 2: Average individual equal error rates (EER). 25-states customer model, 5-state background model, 4 mixtures per states. Both customer and background model are training using $\mathcal{K}$-means followed by 5 EM iterations

EER	Raw Score	Normalized Score
Male	6.49	4.99
Female	7.69	6.62
Average	7.08	5.79

Table 3: Average individual equal error rates (EER). 25-states customer model, 1-state background model, 4 mixtures per states. The customer model is trained using $\mathcal{K}$-means only. The background model is trained using $\mathcal{K}$-means followed by 5 EM iterations

state background model [3].

The raw EER we obtained are different from what Parthasarathy *et al.* obtained on the same database since Parthasarathy *et al.* plugged a speaker independent variance into the customer model λ_c. Another difference is that we used several iterations of the EM algorithm after the $\mathcal{K}$-means training while Parthasarathy *et al.* only used $\mathcal{K}$-means. In a third set of experiments, we reproduced the first experiment without EM iterations. Results are given in table 3. A significant improvement of the raw EER is obtained, while the normalized EER is almost unchanged but still more than 15% better than the raw score. That illustrates the sensitivity of the raw EER to implementation details, and therefore the need for a normalization scheme to improve the system's robustness. However, we should point out that the obtained improvement is still far from the typical 50% of improvement that can be obtained using speaker independent background models on this database.

In a last set of experiments, we investigated the idea of creating a background phrase model by reversing the order of the states after training. A 5 states background model HMM was trained, and after training, the state sequence was inverted. The results are given in Table 4. The results are slightly better that what was obtained using the original state order in Table 2, and confirm that a background model can be designed from the user enrollment data, when an appropriate smearing technique is provided. Further work is needed to define a smoothing technique that would reduce the gap in performance between a speaker dependent and a speaker independent background model.

EER	Raw Score	Normalized Score
Male	6.49	5.29
Female	7.69	6.80
Average	7.08	6.03

Table 4: Average individual equal error rates (EER). 25-states customer model, 5-*reversed* states background model, 4 mixtures per states. The customer model and background model are trained using $\mathcal{K}$-means only.

4. CONCLUSION

State-of-the-art speaker verification systems rely on too many assumptions like for example the availability of speaker independent data or models or of a transcription of the password phrase. These assumptions limit the design of flexible speaker verification applications for portable devices such as palm-top computers, and some new techniques are therefore needed. This paper presents preliminary results towards this goal, where we focus on the design of a speaker verification system where the only data available are the customer enrollment utterances. While our results are still far from the performance level that can be obtained with likelihood ratio scores derived using speaker independent background models, some encouraging improvements have been obtained compared to the use of a simple likelihood score.

5. REFERENCES

[1] C.-H. Lee, F.K. Soong, and K.K. Paliwal, editors. *Automatic Speech and Speaker Recognition: Advanced Topics*, chapter 2. Kluwer Academic Publishers, 1996.

[2] A. L. Higgins, L. Bahler, and J. Porter. Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1:89–106, 1991.

[3] S. Parthasarathy and A. E. Rosenberg. General phrase speaker verification using sub-word background models and likelihood-ratio scoring. In *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, USA, 1996. ICSLP'96.

[4] D. Reynolds. Speaker identification and verification using gaussian mixture models. In *ESCA Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 27–30, 1994.

[5] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong. The use of cohort normalized scores for speaker verification. In *Proc. Int. Conf. on Spoken Language Processing*, pages 599–602, Banff, Alberta, Canada, 1992. ICSLP.

[6] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang. A segmental K-means training procedure for connected word recognition. *AT&T Bell Labs Tech. J.*, 65(3):21–31, 1986.