EXPERIMENTS WITH SPEAKER VERIFICATION OVER THE TELEPHONE

J.L. Gauvain, L.F. Lamel, B. Prouts†

LIMSI - CNRS, B.P. 133, 91403 Orsay, France
{gauvain,lamel}@limsi.fr

†VECSYS, 3 r. de la Terre de Feu - Les Ulis, 91952 Courtabœuf, France

ABSTRACT

In this paper we present a study on speaker verification showing achievable performance levels for both high quality speech and telephone speech and for two operational modes, i.e. text-dependent and text-independent speaker verification. A statistical modeling approach is taken, where for text independent verification the talker is viewed as a source of phones, modeled by a fully connected Markov chain, where the lexical and syntactic structures of the language are approximated by local phonotactic constraints. A first series of experiments were carried out on high quality speech from the BREF corpus to validate this approach and resulted in an *a posteriori* equal error rate of 0.3% in text-dependent as well as in text-independent mode. A second series of experiments were carried out on a telephone corpus recorded specifically for speaker verification algorithm development. On this data, the lowest equal error rate is 2.9% for the text-dependent mode when 2 trials are allowed per attempt and with a minimum of 2s of speech per trial.

INTRODUCTION

Speaker verification has been the subject of active research for many years, and has many potential applications where propriety of information is a concern. With the increasing number of services offered by telephone, accurate verification capability over the telephone could lead to many more near-term applications of this technology. In this paper, we present a study on speaker verification showing achievable performance levels for both high quality speech from the BREF corpus and for telephone speech, in two operational modes, i.e. text-dependent and text-independent verification. The experiments with data from the BREF corpus[9] were carried out to calibrate the algorithm on high quality speech, even though the corpus was not designed to perform speaker recognition experiments. The second speech corpus is a telephone speech corpus especially designed to evaluate speaker recognition algorithms.

A statistical modeling approach is taken, where the talker is viewed as a source of phones, modeled by a fully connected Markov chain[3, 8] and each phone is in turn modeled by a 3 state left-to-right HMM. Verification can be carried out in text-dependent or text-independent mode. For text-dependent verification the model is no longer a fully connected Markov chain as the phone sequence obtained by concatenation of the lexical items is used to constrain the search space. For text-independent verification, the lexical and syntactic structures of the language are approximated by local phonotactic constraints. This approach provides a better model of the talker than can be done with simpler techniques such as long term spectra, VQ codebooks, or a simple Gaussian mixture[16]. The use of small ergodic HMM (with a maximum of 5 to 8 states) has been reported for speaker identification[12, 15, 10]. Gaussian mixture models, which are special cases of ergodic HMM, have been used for speaker identification[13, 16]. The use of phone-based HMM has also been reported for text-dependent[14, 11] and for text-independent, fixed-vocabulary[14] speaker identification.

We have previously applied this phone-based approach to speaker identification[7, 3], where a set of phone models is trained for each speaker. The identification of a speaker from the signal $\mathbf{x}$ is performed by computing the phone-based likelihood $f(\mathbf{x}|\lambda)$ for each speaker λ in the known speaker set. The speaker identity corresponding to the model set with the highest likelihood is then hypothesized. This phone-based approach has been shown to be successful not only for speaker identification but also for gender and language identification[3, 8]. In this paper the same speaker model is applied to speaker verification, and the likelihood ratio $f(\mathbf{x}|\lambda)/f(\mathbf{x})$ is compared to a speaker independent threshold in order to decide acceptance or rejection.

METHODOLOGY

Maximum a posteriori (MAP) estimation is used to generate speaker-specific models from a set of speaker-independent (SI) seed models. The speaker-independent seed models provide estimates of the parameters of the prior densities and also serve as an initial estimate for the segmental MAP algorithm[5]. This approach allows a large number of parameters to be estimated from a small amount of speaker-specific adaptation data.

The technique of phone-based acoustic likelihoods is applied to the problem of speaker-identification as follows. A set of context-independent (CI) phone models are built for each speaker by adaptation of CI, SI seed models using MAP estimation as proposed in[4]. The unknown speech is recognized by all of the speakers' models in parallel, and the hypothisized identity is that associated with the model set having the highest likelihood.

Assuming no prior knowledge about the speaker distribution, the *a posteriori* probability $\Pr(\lambda|\mathbf{x})$ is approximated by the score $L(\mathbf{x};\lambda)$ defined as

$$L(\mathbf{x};\lambda) = f(\mathbf{x}|\lambda)^{\gamma} / \sum_{\lambda'} f(\mathbf{x}|\lambda')^{\gamma}$$

where the λ' are the speaker-specific models for all speakers known to the system and the normalization coefficient γ was empirically determined as 0.02. (This coefficient is needed to compensate for independency approximations in the model.) Calculating the denominator of this expression is very costly as the number of operations is proportional to the number of speakers used in the calculation, or as in our case, the number of target speakers. We can significantly reduce the required computation by using a Viterbi beam search on all the speakers' models in parallel. This decoder, which was developed for speaker identification and the identification of other non-linguistic speech features [3, 8] has been easily modified to provide not only the likelihood of the most probable speaker, $f(\mathbf{x}|\lambda)$, but the likelihoods for the N most probable speakers. We thus reduce the neccessary computation by approximating the summation above by a summation over a short list of the most probable speakers. This implementation is a modified phone recognizer where the output phone string is ignored and only the acoustic likelihood is taken into account.

If a verification attempt is unsuccessful, it is common practice to allow a second trial in order to reduce the false rejection of known users. A straight-forward approach is to base the decision only on the score $L(\mathbf{x};\lambda)$ of the second attempt, ignoring the preceding trial. This approach can be justified on the grounds that the actual test data is potentially invalid. Alternatively, it is possible to base the decision on the scores of both trials.[1] In our system we use this second approach as it reduced the error rate by 21%, compared with an error reduction of 13% obtained using only the score of the last attempt.

EXPERIMENTS WITH BREF

The aim of the first series of experiments was to calibrate the algorithm on high quality speech. For these experiments we made use of a portion of the BREF corpus[9]. This corpus contains read-speech material from 120 speakers, however it was not designed to perform speaker recognition experiments. Since all the data for each speaker was recorded in a single session (lasting about 4 hours), temporal variations in the speakers' voice are likely to be minimum. For each speaker each utterance has a unique prompt text, so it is not possible to assess the use of fixed, identical training and test sentences with this corpus.

For each of 50 target speakers, the first 75 sentences were used as training material. 50 sentences for each speaker were used for verification test. A set of 1820 sentences from 65 of the remaining speakers were used to provide data for impostor attempts.

A set of 35 SI, CI phone models served as seed models for estimation of speaker-specific models for each target speaker. Only the Gaussian means were adapted, i.e., the estimates of the variances are the same for all speakers. A common silence model was used for all speakers. A feature vector containing 16 Mel-frequency scale cepstrum coefficients (8kHz bandwidth) and their first and second order derivatives was computed every 10 ms. Cepstral-mean removal was performed for each sentence.

Condition	*EER*
Gaussian mixture, 3s	5.5%
Text unknown, 3s	1.1%
Text known, 3s	1.0%
Text known, 5s	0.7%
Text unknown, EOS (avg 7.1s)	0.3%
Text known, EOS (avg 7.1s)	0.3%
Text known, EOS (avg 7.1s), 2 trials	0.2%

Table 1: *A posteriori* equal error rates (EER) for different model sets and operational modes on a total of 4370 attempts for users and impostors with 1 trial per attempt. The amount of speech data used for verification is specified: "3s" corresponds to the first 3s of each utterance; "EOS" means the entire sentence used with an average duration of 7.1s per sentence.

Speaker verification performance using phone-based models was compared to a simpler system based on long-term statistics of the speech, using Gaussian mixture. In this case, the silence is modeled using a single mixture of 32 Gaussians which is common for all speakers. The speech portion of the training data for each speaker is modeled using a single mixture of 32 Gaussians. With this Gaussian mixture model an *a posteriori* equal error rate (EER) of 1.8% was obtained on a set of 50 test sentences per speaker recorded immediately after the training material, using 3s of speech per trial and one validation trial. On a second set of 50 test sentences taken from the end of each recording session the *a posteriori* EER was 5.5%. Therefore this second set of test sentences were used in all the remaining experiments in order to have more realistic conditions.

The experimental results on the BREF data are summarized in Table 1. This approach was evaluated in both text-independent and text-dependent modes, for one and two trials per validation attempt. When 2 verification trials are authorized (for target speakers and impostors), there are on average 1.1 trials per attempt. All conditions are compared using the *a posteriori* EER.

With a fixed amount of 3s of speech per trial and one trial per validation attempt, the EER is 1.1% in text-independent mode (text unknown) for a total of 4370 attempts (by users and impostors). If the text is known the EER is reduced to 1.0%. Using a fixed longer duration of 5s of speech, the EER is 0.7%. If the entire utterance is used for the verification, the EER with a maximum of two trials per attempt is 0.2%. It should be noted that this number is certainly optimistic in that the training and verification data were recorded in the same 4h session.

In order to assess the performance with a smaller amount of training data for each speaker, the same experiments were carried out using only 30 and 10 sentences to estimate the speaker specific phone models. Reducing the available training data in half (i.e. 30 sentences instead of 75) did not affect the performance: the EER with 1 trial in text dependent mode and 3s of speech per trial still being 1.0%. For the same test conditions the verification EER is 1.8%, when only 10 sentences per speaker were used for training. With these same models, using the entire utterance sentence for test and a maximum of 2 trials per attempt, resulted in an EER of 0.4%.

[1] It is evidently possible to allow more than 2 trials per attempt, in which case the score would take into account all scores from previous trials.

EXPERIMENTS WITH TELEPHONE SPEECH

In order to carry out experiments on speaker verification over dialed-up telephone lines, a corpus has been designed specifically for this purpose. The corpus contains training material from 100 target speakers, recorded in multiple sessions with variable of telephone handsets and calling locations. Each target speaker is also providing multiple sessions of test data to be used for verification attempts. Data from 1000 unknown speakers is being collected so as to provided test material from impostors. The recordings are similar to the Polyphone recordings being collected in several languages[2, 1], with different types of read speech material (words, numbers, dates, phonetically compact sentences, ellicited and spontaneous speech, etc.) so as to be able to assess the effects of data type on the verification accuracy.

In these experiments only a portion of the corpus is used. Approximately 75 sentences from each of 45 target speakers (coming from 2 recording sessions) were used to adapt SI, CI seed models to form speaker-specific models. As for the experiments with BREF only the Gaussian means were adapted, and a common silence model was used for all speakers. A feature vector containing 13 Mel-frequency scale cepstrum coefficients (0-3.5kHz bandwidth) and their first order derivatives was computed every 10 ms. In order to minimize effects due to channel differences, cepstral-mean removal was performed for each sentence. The impostor data include 1980 sentences from 135 speakers, with each speaker participating in one call.

Figure 1 gives ROC (Receiver Operating Characteristics) curves for different model types and operational modes for the telephone data. The phone-based approach is compared to a baseline system using Gaussian mixture. Two mixtures of 32 Gaussians are used, one for silence/noise (common for all speakers) and another for the speech, specific to each speaker. For the phone-based approach, text-dependent and text-independent modes are compared, for one and two verification trials.

The ROC curve for the Gaussian mixture model is shown in (a). This can be compared with (b) the ROC of the phone-based approach in text-independent mode. The phone-based approach is seen to perform significantly better than the Gaussian mixture model (7.3% v.s 9.0% EER) with 1 only trial per attempt and an average of 3.2s of speech per trial. If the text is known, the EER is reduced to 5.1% (curve c). It should be noted that with the phone-based approach, knowing the text does not imply the use of a fixed text. The user can be prompted to read any text. In (d), 2 verification trials are allowed per attempt, reducing the EER to 4.1% with 1.1 trials on average. Curve (e) shows the ROC if a minimum amount of 2s of speech is required for each trial. For the sentences having this minimal duration, the EER is reduced to 2.9%.

From these ROC curves we can make the following conclusions:

- The phone-based approach performs better than the simpler approach based on a mixture of Gaussians.

- On the telephone corpus, better results are obtained when the text is known *a priori*. It should be kept in mind that these experiments were carried out with an orthographic transcription of the speech, and not with the prompt text. The same experiments should be carried out with the prompt text to more realistically estimate the performance.

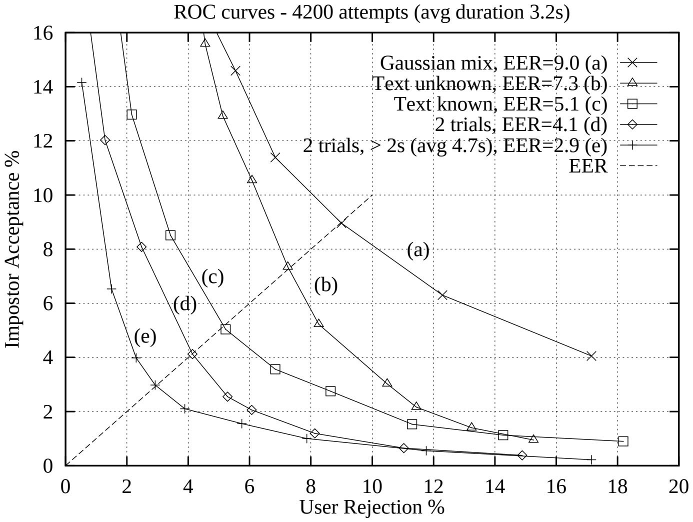

Figure 1: ROC curves for different model sets and operational modes: (a) multi-Gaussian model; (b) 35 phone models, text independent; (c) 35 phone models, text dependent; (d) same as (c) with 2 trials; (e) same (d) with at least 2s of speech. The dotted line shows the points of equal error (false acceptance/false rejection).

- Allowing a second verification trial reduces the EER without significantly increasing the number of trials for the target speakers (10% increase).

- Requiring a minimum speech signal duration of 2s reduces the error rate by almost 30%.

The cepstral analysis used to represent the speech signal essentially models the spectral envelope. However, it is widely believed that the fundamental frequency (F0) contains information about the identity of the speaker, which may be complementary to the information captured by the cepstral analysis. To investigate this possibility, the F0 was estimated for all the frames of the training and test data. A Gaussian model was built for the log F0 of each speaker. Using only this model, an EER of 19.3% was obtained with one verification trial, and an EER of 17.3% when two trials were allowed with an average of 1.3 trials per target speaker attempt. When both the cepstral and F0 models were used, the EER was reduced by 7% with one trial compared to the cepstral-based phone model only, but no significant improvement was observed when 2 trials were authorized.

ERROR ANALYSIS

For the telephone speech data, the EER for the best configuration tested (minimum duration of 2s and 2 trials) is 2.9%. To understand this relatively high error rate, histograms of the scores for target speakers and impostors (2200 attempts for each class) are shown in Figures 2a and 2b. In Figure 2a the distributions appear to be well separated: 87% of the attempts by impostors have a score of essentially 0, and 51% of the attempts by target speakers

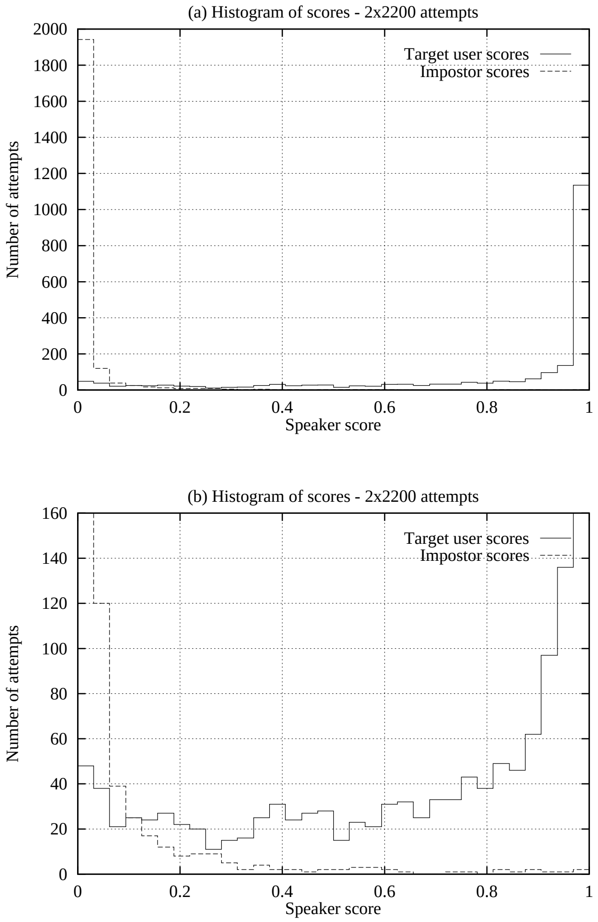

Figure 2: Distribution of scores for target speakers and impostors.

have a score of essentially 1. The overlap in the distributions is better seen in Figure 2b with an expanded vertical scale. It is apparent in these histograms that main source of error is the low score for certain attempts by target speakers. Almost 2% of the attempts by target speakers have a score almost equal to 0.

The errors are localized on a few speakers for whom the verification error is high. Attempts by 5 of the target speakers account for 70% of all errors. For these 5 speakers the verification error rate ranges from 36% to 70%. A large proportion of the errors seem to be due to variability in the origin of the calls, the channel conditions or large level variations for a given target speaker, and not characteristics of the speaker.

SUMMARY

We have presented a series of experiments in speaker verification for both high quality speech and telephone speech using a statistical approach based on HMM phone models. The decoding procedure has been efficiently implemented by processing all the models in parallel using a time-synchronous beam search strategy. Speaker verification (or identification) can be carried out in both text-dependent or text-independent modes using the same phone models.

For text-independent verification, the phone based approach was shown to clearly out-perform a simpler Gaussian mixture model on high-quality speech from the BREF corpus and for telephone speech has a 20% lower *a posteriori* equal error rate. For the BREF corpus, text-independent and text-dependent verification EERs were about the same.[2] On the telephone corpus, text-dependent verification performs better than text-independent. When a verification attempt fails, allowing a second trial reduces the number of errors by 20%, while only increasing the number of trials by 10%. For the telephone speech corpus, the majority of the errors are due to low scores for a few target speakers, mostly reflecting differences in the origin of the call for the training and testing sessions. In an additional experiment on the telephone speech, combining a model for F0 with the speaker-specific phone model set did not significantly improve performance. On the telephone speech corpus, an *a posteriori* equal error rate of 2.9% was obtained using a minimum duration of 2s per trial, in text-dependent mode, allowing 2 trials per attempt. This can be contrasted with the equal error rate obtained on the high quality speech corpus which is well under 1%.

REFERENCES

[1] J. Bernstein, K. Taussig, J. Godfrey, “MACROPHONE: An American English Telephone Speech Corpus for the Polyphone Project,” *ICASSP-94*.

[2] J. Godfrey, “Multilingual Speech Databases at LDC,” *ARPA Human Language Technology Workshop*, Plainsboro, NJ, March 1994.

[3] J.L. Gauvain, L. Lamel, “Identification of Non-Linguistic Speech Features,” *ARPA Human Language Technology Workshop*, Plainsboro, NJ, March 1993.

[4] J.L. Gauvain, C.H. Lee, “Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities,” *Speech Communication*, **11**(2-3), June 1992.

[5] J.L. Gauvain, C.H. Lee, “Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” *IEEE Trans. on Speech & Audio*, **2**(2), April 1994.

[6] H. Gish , M. Schmidt, “Text-Independent Speaker Identification,” *IEEE Signal Processing Magazine*, **11**(4), Oct. 1994.

[7] L. Lamel, J.L. Gauvain, “Continuous Speech Recognition at LIMSI,” *Final review of the DARPA Artificial Neural Network Technology Speech Program*, Stanford, CA, Sept. 1992.

[8] L. Lamel, J.L. Gauvain, “A Phone-based Approach to Non-Linguistic Speech Feature Identification,” *Computer Speech &Language*, **9**(1), Jan. 1995.

[9] L. Lamel, J.L. Gauvain, M. Eskénazi, “BREF, a Large Vocabulary Spoken Corpus for French,” *Eurospeech’91*.

[10] T. Matsui, S. Furui, “Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMs,” *ICASSP-92*.

[11] T. Matsui, S. Furui, “Concatenated Phoneme Models for Text-Variable Speaker Recognition,” *ICASSP-93*.

[12] A. Poritz, “Linear Predictive Hidden Markov Models and the Speech Signal,” *ICASSP-82*.

[13] R. Rose , D. Reynolds, “Text Independent Speaker Identification using Automatic Acoustic Segmentation,” *ICASSP-90*.

[14] A. Rosenberg, C.H. Lee, F. Soong, “Sub-Word Unit Talker Verification Using Hidden Markov Models,” *ICASSP-90*.

[15] N. Tishby, “On the Application of Mixture AR Hidden Markov Models to Text-Independent Speaker Recognition,” *IEEE Trans. Signal Processing*, **39**,(3), March 1991.

[16] B. Tseng, F. Soong, A. Rosenberg, “Continuous Probabilistic Acoustic MAP for Speaker Recognition,” *ICASSP-92*.

[2] This has been previously observed for speaker identification on the TIMIT corpus[8].