SPEAKER RECOGNITION WITH THE SWITCHBOARD CORPUS

Lori Lamel and Jean-Luc Gauvain

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{lamel,gauvain}@limsi.fr

ABSTRACT

In this paper we present our development work carried out in preparation for the March'96 speaker recognition test on the Switchboard corpus organized by NIST. The speaker verification system evaluated was a Gaussian mixture model. We provide experimental results on the development test and evaluation test data, and some experiments carried out since the evaluation comparing the GMM with a phone-based approach. Better performance is obtained by training on data from multiple sessions, and with different handsets. High error rates are obtained even using a phone-based approach both with and without the use of orthographic transcriptions of the training data. We also describe a human perceptual test carried out on a subset of the development data, which demonstrates the difficulty human listeners had with this task.

INTRODUCTION

This paper describes a speaker verification system evaluated in the March'96 speaker recognition test organized by NIST [6], our development work carried out in preparation for the test, and further experiments we have carried out since the evaluation. This was our first participation in a speaker recognition evaluation, and allowed us to compare our approach to the approaches used by the 8 other sites on a common corpus. Published results on speaker recognition typically either make use of proprietary corpora specifically designed for the task, or widely-used corpora, such as TIMIT, that were not designed for such purposes and have the default that all the data for a speaker was recorded in a single session. The Switchboard corpus, while not ideal for speaker recognition, contains data for speakers recorded in multipe sessions (different calls) and from different locations (different handsets).

The usual approach taken at LIMSI to speaker recognition makes use of a phone-based speaker model [4], where the talker is viewed as a source of phones, modeled by a fully connected Markov chain. Each phone is 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 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 can be 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.

We have previously applied this phone-based approach to speaker identification[4] and speaker verification [3]. 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 [4]. Applying the approach to speaker verification, the likelihood ratio $f(\mathbf{x}|\lambda)/f(\mathbf{x})$ is compared to a speaker-independent threshold in order to decide acceptance or rejection [3].

CORPUS AND EXPERIMENTAL CONDITIONS

The March'96 speaker evaluation compared performance for 3 training conditions and 3 test durations[1]. While all 3 training conditions contained 2 minutes of speech per speaker, they contrasted training with a *single-session* (two minutes of speech taken from the same call), with *single-handset* (one minute of data from each of two calls using the same telephone), and with *two-handsets* (one minute of speech from two different calls using different telephones). The speaker recognition performance was measured for test segments of 3s, 10s, and 30s. No transcriptions of the training data were provided, and only unsupervised training techniques were permitted. Since Switchboard is a corpus of telephone conversations, the training and test data were actually formed of concatenated segments corresponding to one side of the conversation, with silences removed. While this eliminated the problem of crosstalk, the resulting speech is choppy and a bit strange without the silences that normally appear in conversations.

The development set contained training and test data from 88 speakers (43 male/45 female). There were a total of 402 test segments for each duration. The evaluation set contained training and test data from 19 female speakers and 21 male speakers. The test data included samples from same-sex and cross-sex impostors. For each test segment duration there were about total of 1300 speech samples from target speakers. For the 3s and 30s durations there were about 1100 trials from impostors, roughly 50% of each sex. For the 10s test segments there were about 500 trials from impostors of the same sex.[1]

For development trials, the likelihood computation is carried out in parallel for all models time synchronously, and models for which the partial likelihood at time t is significantly lower than the best model are discarded. Only the N highest likelihoods are used to normalize the speaker score. For the evaluation test, each test-sample target-model pair is run independently. For normalization scores for all of the 88 reference speakers in the development set are obtained by setting the pruning threshold to infinity.

SYSTEM DEVELOPMENT AND EXPERIMENTAL RESULTS

System development was carried out using the development training data and the 10s test segments. Although the evaluation was for speaker verification (i.e., a yes/no decision), development was carried out using the closed-set speaker identification rate as a performance measure, as we have observed this to be a close indicator of verification performance.

In Table 1 we give the closed-set speaker identification rates on the 10s devtest data for different training configurations, for a Gaussian mixture model (GMM) and two sets of phone models (with 46 and 12 phones). For the GMM we compared the use of different training data (single session s1a, single handset s1a+s2, two handset s1a+hs2), different models (32 to 256 Gaussians, 1 or 2 mixtures), and different analyses (MFCC and PLP). The acoustic feature set contains 12 MFCCs and their first and second order derivatives, and the log energy and it's first and second order derivatives, computed every 10ms on a 30ms window. Cepstral mean removal is performed on each training and test sample. The highest speaker-identification (SID) rate is obtained with the two-handset training condition and a single Gaussian mixture with 128 Gaussians. No difference in performance was observed with PLP and MFCC.

For the phone-based approach, the s1a+hs2 training condition was used with an MFCC analysis. Maximum a posteriori (MAP) estimation is used to generate speaker-specific models from a set of speaker-independent (SI) seed models trained on the Macrophone corpus. These seed models provide estimates of the parameters of the prior densities and also serve as an initial estimate for the segmental MAP algorithm[2]. This approach allows a large number of parameters to be estimated from a small amount of speaker-specific adaptation data. For unsupervised adaptation, the training data is first labeled using the SI models. Since the training data is relatively limited, performance using a reduced set of 12 phone models was compared to that with the original 46 phone models. In this case the labeled segments were mapped to the 12 phone set. In Table 1 the SID rates are slightly higher for the phone-based approach than for the GMM. If the Switchboard orthographic transcriptions are used (ie. supervised model adaptation), there is an additional small improvement in SID rate. These results are surprising as in the past we found the phone-based approach to outperform more simple models such as Gaussian mixture on the BREF corpus [5] and the VECLIM telephone speech corpus (100 targets, 1000 impostors) [3], specifically designed to carry out speaker recognition experiments. We attribute the inability of the phone-based approach to outperform the GMM to the need for more accurate phone models (better transcriptions) and/or a mismatch in training corpora. These transcriptions have in the past been obtained either from the orthographic transcription (not allowed in the test conditions) or have been automatically generated using speaker-independent phone models. The phone accuracies obtained on the other corpora are relatively high (BREF: 87%, TIMIT: 75%), while for Switchboard the phone accuracy is only about 35% (50% when rescored with 12 phones).

Training condition	*SID rate*
GMM s1a (32g)	64.2%
GMM s1a (64g)	66.9%
GMM s1a (2x32g)	62.4%
GMM s1a+s1b (64g)	72.2%
GMM s1a+s2 (64g)	79.1%
GMM s1a+hs2 (64g)	83.1%
GMM s1a+hs2 (2x64g)	71.6%
GMM s1a+hs2 (128g)	84.3%*
GMM s1a+hs2 (128g, PLP)	84.3%
GMM s1a+hs2 (256g)	83.3%
46 phones, bg, unsupervised	86.6%
12 phones, bg, unsupervised	86.8%
46 phones, bg, transcriptions	87.1%
12 phones, bg, transcriptions	88.1%

Table 1: Training conditions and SID rates on March96 10s devtest data. *corresponds to the evaluation system

Figure 1 shows the commonly used ROC curves and the the detection cost function (DCF)[1] curves for the development test data with the GMM.[2] The DCF is relatively constant over a range of decision thresholds, thus the exact value is not crucial to be near the minimum.

Table 2 gives the speaker identification rates and DCF values obtained on the devtest data for the GMM and the

[1] No data from impostors (non-target speakers) was provided for system development, therefore results on the dev data are reported using simulated impostors.

[2] The DCF used for this evaluation is:

$$DCF = 0.1 * P_{Miss|Target} + 0.99 * P_{FalseAlarm|NonTarget}$$

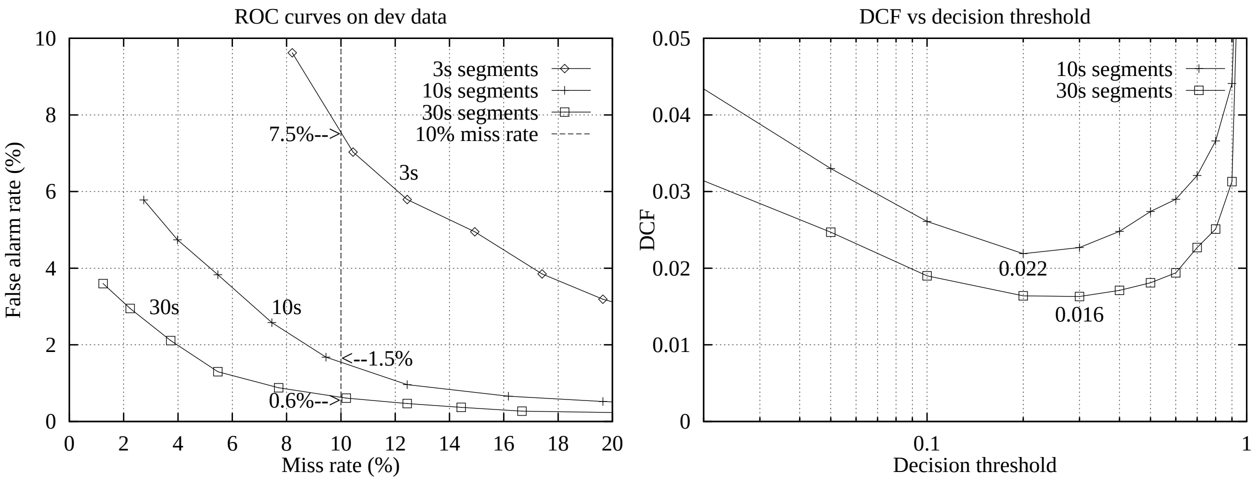

Figure 1: ROC curves and DCF for GMM on devtest data. 88 target speakers, 402 target trials, 34,974 non-target trials (simulated impostors).

Test	*SID rate(%) / DCF*		
Configuration	3s	10s	30s
GMM	69.2 / .041	84.3 / .022	89.1 / .016
12p,unsup	71.4 / .038	86.8 / .022	92.8 / .008
46p,unsup	73.1 / .036	86.6 / .021	91.8 / .009
12p,trans	72.1 / .038	88.1 / .021	91.8 / .008
46p,trans	72.4 / .036	87.1 / .021	92.3 / .009

Table 2: SID rates and minimum DCF on devtest data for GMM and phone-based models.

Test	*SID rate(%) / DCF*		
Configuration	3s	10s	30s
GMM f	50.7 / .063	68.5 / .048	75.4 / .033
GMM m	50.7 / .066	66.8 / .050	73.6 / .036
GMM f+m	50.7 / .065	67.7 / .049	74.5 / .034

Table 3: SID rates and DCF for the GMM on the eval test data.

phone-based models, with and without the use of training transcriptions. The use of transcriptions in training improves the speaker identification rate for the 3s and 10s segments with 12 phone models, yet the performance is slightly worse for the 30s segments. With 46 phone models, the speaker identification performance improves only for 10s and 30s segments. In all cases the DCF remains the same.

At the time of the March'96 evaluation we only had complete development results for the 10s test segments.[3] Therefore, in consideration of the reduced computational requirements, and the lack of a clear advantage for the phone-based approach, we decided to use a GMM for the speaker recognition evaluation. The training configuration was that which had the highest speaker identification rate on the development data, that is, a 128 component Gaussian mixture estimated on the two-handset training data. The SID rates and DCF are given for the evaltest data in Table 3. The speaker identification rates are substantially lower then those obtained on the devtest data, and the DCF values are higher.

HUMAN PERFORMANCE

Errors with the GMM are relatively diffuse, which is in contrast to our experience with the VECLIM corpus[3] where the errors tended to be focused on a few speakers. To investigate further the data, a perceptual experiment was carried out to assess the ability of humans to discriminate speakers on the Switchboard data [8]. Test tokens were selected from twenty speakers (8 male, 12 female) who were often confused (4 to 15 times) during automatic verification. Tokens from an additional 24 speakers (12 male, 12 female) with which the reference speakers were confused served as impostor data. Two hundred sample pairs were constructed from the same-speaker/same-conversation, same-speaker/different-conversation, same-speaker/different-handset, and different-speakers. The reference speaker samples were taken from the development training data, and the test speaker samples were taken from both the training data and the test data.

Eight listeners participated in the AX listening task, indicating "same speaker" or "different speaker" on a 3-scale confidence rating (+,0,–). Results are given in Table 4. Human subjects had the most difficulty identifying data samples from the same speaker using different handsets, followed by samples from different conversations. These factors also are known to cause problems for automatic systems. In contrast to the system evaluation where a false acceptance was considered 10 times worse than a false rejection, listeners were not given instructions with respect to the two types of errors. All but listener 2 had more false rejections than false acceptations, and listeners 5 and 7 were clearly preferred to reject if uncertain.

[3] For the 30s test data the DCF for the phone-based approach is half that of the GMM (see Table 2. This suggests that the phone-based approach needs longer speech segments to outperform the GMM.

Conditions	*Test Subjects*								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Average*
Diff hands (20)	50	40	45	60	80	50	75	65	58
Diff conv (20)	25	10	25	35	35	20	40	30	28
Same conv (20)	5	0	0	20	20	5	10	0	8
No info (5)	40	40	40	40	40	20	60	60	43
Diff spkr (135)	21	27	14	10	2	16	2	13	13
False rejections	28	18	25	38	45	27	43	34	32
False acceptances	21	27	14	10	2	16	2	13	13
Total error rate	24	24	18	20	16	19	16	20	19

Table 4: Human capability of classifying voices as "same" or "different" as a function of speech condition, given in percentage error. Test subjects 1-5 are non-native English speakers, test subjects 6-8 are native English speakers. The number of test trials for each condition is given in parentheses.

As shown in Table 5 humans had an error rate of about 40% when they were unsure compared to under 10% when they were confident. Listeners 5 and 7 never made an error when they claimed to be sure. 30 of the 56 tied-pairs corresponding to confusions with the GMM system were confused by humans, and over half of the confusions made by humans were also made by the system. Automatic system performance on the test tokens in common with the perceptual experiment is on the order of 10%, which is better than the total error rate for humans on the 200 token pairs.

Conf. rating	*Error rate*	*Conf. rating*	*Error rate*
same +	8	different +	10
same 0	30	different 0	15
same –	46	different –	37

Table 5: Confidence ratings of the 8 test subjects for "same" and "different" judgements (+ very sure, 0 sure, – not sure).

CONCLUSIONS

In this paper we have described our development work in preparation for the March'96 Speaker Recognition evaluation organized by NIST. We found that on the Switchboard data, a phone-based approach to speaker identification did not perform substantially better than a more simple Gaussian mixture model. While this is in contrast to our previous observations on both high quality and telephone based corpora[4, 3], the same effect was observed by other participants in the evaluation. Using the GMM approach, the best verification results were obtained when a single mixture of 128 Gaussians was trained with data from different conversations and different telephone handsets. We attributed the inability of the phone-based approach in unsupervised mode to outperform the GMM to the need for more accurate phone models. We thus investigated the performance using the orthographic transcriptions of the 2 min training data. While for some conditions slight improvements were obtained, the gain was less than anticipated. This is likely to be due to a variety of factors including the unnaturalness of the concatenated training segments and the resulting concatenated transcriptions, the full word form of the transcriptions that often do not reflect the reduced forms found in conversational speech, the use of SI seed models trained on a portion of read sentences from the Macrophone corpus.

A perceptual test was carried out to assess speaker verification human performance. The test subjects had difficulty in classifying a pair of tokens from the same speaker when the tokens came from different conversations (28% false rejection) or different handsets (58% false rejection). Subjects tended to reject more easily than to accept, and native English speakers performed better than non-native. The task was difficult for humans, and even when the subjects expressed confidence in their response they still had an error rate of almost 10%.

ACKNOWLEDGEMENT

The authors wish to acknowledge the participation of Simo Goddijn, from Utrecht University who carried out the perceptual experiments during an Erasmus exchange internship.

REFERENCES

[1] G. Doddington, "The 1996 Speaker Recognition Evaluation Plan," Jan. 1996.

[2] 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.

[3] J.L. Gauvain, L.F. Lamel, B. Prouts, "Experiments with speaker verification over the telephone," *ESCA Eurospeech'95*.

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

[5] L. Lamel, J.L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," *ESCA Eurospeech'91*.

[6] A. Martin, M. Przybocki, J. Fiscus, D. Pallett, "Evaluation on Switchboard Corpus Selected Segments," presented at the *NIST Speaker Recognition Workshop*, Linthicum, March 1996.

[7] L.F. Lamel, J.L. Gauvain, "LIMSI March'96 Speaker Recognition Evaluation," presented at the *NIST Speaker Recognition Workshop*, Linthicum, March 1996.

[8] S. Goddijn, "Testing the LIMSI-Algorithm for Speaker Verification on two Different Corpora," Erasmus Project Report, Utrecht University, The Netherlands, May 1996.