The CLEAR'06 LIMSI Acoustic Speaker Identification System for CHIL Seminars*

Claude Barras, Xuan Zhu, Jean-Luc Gauvain, and Lori Lamel

Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay cedex, France
{barras,xuan,gauvain,lamel}@limsi.fr

Abstract. This paper summarizes the LIMSI participation in the CLEAR'06 acoustic speaker identification task that aims to identify speakers in CHIL seminars via the acoustic channel. The system consists of a standard Gaussian mixture model based system similar to systems developed for the NIST speaker recognition evaluations and includies feature warping of cepstral coefficients and MAP adaptation of a Universal Background Model. Several computational optimizations were implemented for real-time efficiency: stochastic frame subsampling for training, top-Gaussians scoring and auto-adaptive pruning for the tests, speeding up the system by more than a factor of ten.

1 Introduction

The European Integrated Project CHIL[1] is exploring new paradigms for human-computer interaction and developing user interfaces which can track and identify people and take appropriate actions based on the context. One of the CHIL services aims to provide support for lecture and meeting situations, and automatic person identification is obviously a key feature of smart rooms. CHIL has supported the CLEAR'06 evaluation, where audio, video and multi-modal person identification tasks were evaluated in the context of CHIL seminars. Our work at LIMSI focuses on the acoustic modality. The CLEAR'06 acoustic speaker identification task is a text-independent, closed-set identification task with far-field microphone array training and test conditions. Enrollment data of 15 and 30 seconds are provided for the 26 target speakers and test segment durations of 1, 5 10 and 20 seconds are considered [5].

This paper describes the LIMSI acoustic speaker identification system, evaluated in the CLEAR'06 benchmark. The system is a standard GMM-UBM system based on technology developed for use in NIST speaker recognition evaluations. In the next section, the LIMSI speaker recognition system is presented along with specific computation optimizations that were developed for this system. Section 3 gives experimental results on the CLEAR development data and evaluation data.

* This work was partially financed by the European Commission under the FP6 Integrated Project IP 506909 Chil.

[1] CHIL – Computers in the Human Interaction Loop, http://chil.server.de/

R. Stiefelhagen and J. Garofolo (Eds.): CLEAR 2006, LNCS 4122, pp. 233–240, 2007.

2 Speaker Recognition System

In this section, the LIMSI speaker recognition system and several computational optimizations that were implemented for real-time efficiency are described.

2.1 Front-End

Acoustic features are extracted from the speech signal every 10ms using a 30ms window. The feature vector consists of 15 PLP-like cepstrum coefficients computed on a Mel frequency scale, their Δ and Δ-Δ coefficients plus the Δ and Δ-Δ log-energy for a total of 47 features. Ten percent of the frames with the lowest energy are filtered out, on the assumption that they carry less information characteristic of the speaker. No speech activity detection (SAD) module is used in this configuration since silences longer than one second according to the reference transcriptions are a priori removed from evaluation data.

Feature warping [6] is then performed over a sliding window of 3 seconds, in order to map the cepstral feature distribution to a normal distribution and reduce the non-stationary effects of the acoustic environment. In the NIST speaker recognition evaluations, feature warping was shown to outperform the standard cepstral mean substraction (CMS) approach [1].

2.2 Models and Identification

A Gaussian mixture-model (GMM) with diagonal covariance matrices is used as a gender-independent Universal Background Model (UBM). For each target speaker, a speaker-specific GMM is trained by Maximum A Posteriori (MAP) adaptation [3] of the Gaussian means of the UBM. The GMM-UBM approach has proved to be very successful for text-independent speaker recognition, since it allows the robust estimation of the target models even with a limited amount of enrollment data [7]. During the identification phase, each test segment X is scored against all targets λ_k in parallel and the target model with the highest log-likelihood is chosen: $k^* = \operatorname{argmax}_k \log f(X|\lambda_k)$.

2.3 Optimizations

In the CHIL framework, target model training and speaker identification need to be performed efficiently, in faster than real-time for realistic configurations. Several

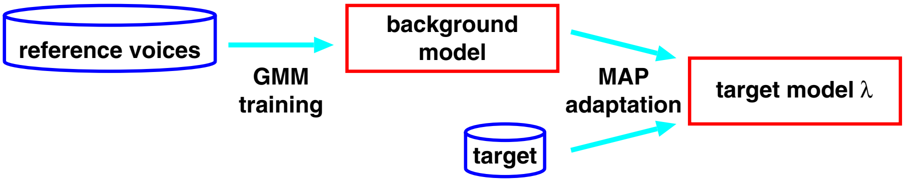

Fig. 1. MAP adaptation of background model to a target speaker

optimizations have thus been implemented addressing training and scoring computational requirements.

Stochastic Frame Subsampling. For speaker recognition, the reduction of the number of frame by a decimation factor up to 10 on the test segment only results in a limited loss of accuracy [4]. This can be explained by the high correlation of neighboring frames and the fact that a temporal context of several frames is already taken into account by the delta and delta-delta coefficients. It can be also of interest to speed up the training of the models. The UBM needs to account for the largest possible speaker variability in the acoustic context of the application; but the amount of training data needs to be put in relation with the number of parameters in the UBM. For training a GMM with diagonal covariance matrices, a few hundred frames per Gaussian should be enough for a reliable estimation of the means and variances. A possible solution can be a fixed rate subsampling as described above; in this situation, a subset of the frames is selected once for all. We have experimented with another schema. For each Expectation-Maximization (EM) iteration of the GMM reestimation, a random selection of frames is applied according to a target ratio. This way, each frame can possibly impact the training. Also, if we train the GMM using a splitting algorithm starting with a single Gaussian, the stochastic frames sampling dramatically speeds up the initial training phases by adapting the number of frames to the number of components.

Top-Gaussian Scoring. The top-Gaussian scoring is an optimization used for speaker verification in the context of the parallel scoring of a set of target models MAP-adapted from the same GMM-UBM [4]. For each frame, the top scoring components of the UBM are selected; then the log-likelihood estimation for all target models is restricted to the same set of components. The speedup increases along with the size of the models and with the number of target speakers.

Auto-Adaptive Pruning. During scoring, it is usual to exclude models with a too low likelihood relative to the best current hypothesis. However in the context of top-Gaussian scoring, the computation is dominated by the UBM initial likelihood estimation and a reduction in the number of target candidates only provides a minor improvement; the major gain is observed when a single model remains and the end of the test segment can thus be discarded. Taking an early decision about the current speaker is also of interest in the context of an online system as required for some CHIL applications. In this situation, an a priori fixed threshold is not precise enough for such an aggressive pruning because of the acoustic variability. We have thus implemented an auto-adaptive pruning, which takes into account the distribution of the best hypothesis log-likelihood:

- at each frame x_t, for each model λ_k, compute its cumulated log-likelihood: $l_k(t) = \frac{1}{t} \log f(x_1 \dots x_t|\lambda_k)$
- choose the best cumulated score up to the current frame: $l^*(t) = \max_k l_k(t)$
- compute the statistics $(\mu_l(t),\sigma_l(t))$ of $l^*(t)$ with an exponential decay factor $\alpha \in]0;1]$ in order to focus on the most recent acoustic context:

$$\mu_l(t) = \frac{1}{\sum_{i=0}^{t} \alpha^i} \sum_{i=0}^{t} \alpha^i l^*(t-i) \text{ and } \sigma_l(t)^2 = \frac{1}{\sum_{i=0}^{t} \alpha^i} \sum_{i=0}^{t} \alpha^i l^{*2}(t-i) - \mu_l(t)^2$$

- initialize $l^*(t)$ on a minimal count $d_{min}$of a few tens to a few hundreds frames
- during scoring, cut model λ_k if $l_k(t) < \mu_l(t) - \lambda(t)\sigma_l(t)$ with the standard deviation factor $\lambda(t)$ either constant or decreasing in time.

3 Experiments

In this section the experimental conditions are described, and the impact of the optimization and development work using the CHIL'05 evaluation data are given. Results on the CLEAR'06 evaluation data are also provided.

3.1 Experimental Setup

Seminars recorded for the CHIL project were used for building the system. All processing were performed on 16 kHz, 16 bits single channel audio files in far-field microphone condition. CHIL jun'04 data (28 segments from 7 seminars recorded by UKA for a total of 140 min.) and dev'06 data (another 140 min. from UKA plus 45 min. from AIT, IBM and UPC partners) were used for training a generic speaker model. Beamformed data were supplied by our CHIL partner ISL/UKA for both the jun'04 and dev'06 data sets. The data from CHIL 2005 speaker identification evaluation (jan'05) was used for the development of the system. For CLEAR'06 evaluation data, the 64 channels of a MarkIII microphone array were provided. However, only the 4th channel of the MarkIII microphone array as extracted and downsampled to 16kHz by ELDA was used.

A gender-independent UBM with 256 Gaussians was trained on speech extracted from jun'04 and dev'06 CHIL data. The amount of data was limited to 2 min. per speaker in order to increase the speaker variability in the UBM, for a total duration of about 90 min. Target models were MAP-adapted using 3 iterations of the EM algorithm and a prior factor of 10. Computation times were estimated on a standard desktop PC/Linux with a 3GHz Pentium 4 CPU and are expressed in Real-Time factor (xRT) when relevant.

3.2 Optimization Results

The effect of the stochastic frame subsampling was studied on the 90 min. of training data, which account for $d \simeq 500.000$ frames after filtering of low-energy frames. With $M = 256$ components in the GMM and $f = 200$ frames kept in average per Gaussian, the gain relative to the standard training using all the frames at each step of the EM estimation is: $g(f) = \frac{d}{M*f} = 500.000/(256*200) \approx 10$. Figure 2 shows the likelihood of the UBM on the training data as a function of the computation time for the stochastic subsampling with an average count of 200 frames per Gaussian, compared to the standard training and to a fixed-rate subsampling with the corresponding 10% ratio; it was obtained by varying the number of EM iterations from 1 to 9. For a given computation time, the stochastic subsampling outperforms the standard training, and also the fixed-rate decimation, due to the faster initialization procedure. For a given EM iteration count, we also observed that the stochastic subsampling even outperforms the full training up to 5 EM iterations, and the fixed-rate subsampling in all configurations.

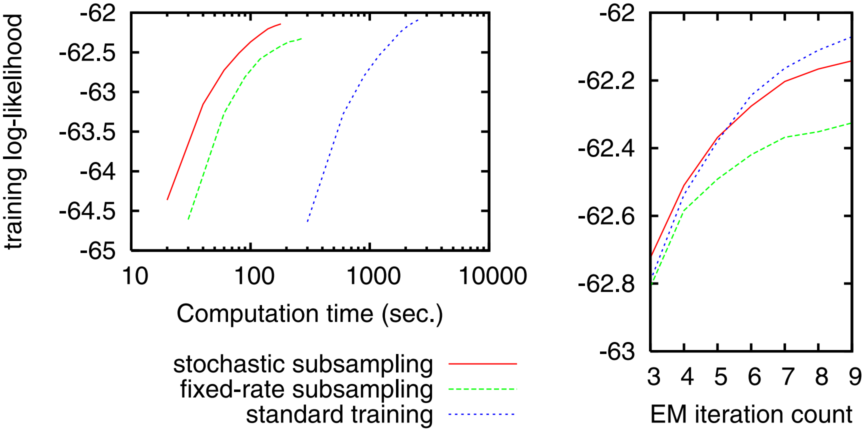

Fig. 2. Likelihood of UBM on training data as a function of computation time and of EM iteration count for standard training, stochastic subsampling and fixed-rate subsampling

The scoring was performed with the top Gaussians. With $M = 256$ components in the GMMs, $T = 10$ top components and $N = 26$ target models, the gain in computation is $g(T) = \frac{M*N}{M+T*N} = (256*26)/(256+10*26) \approx 13$. The pruning with $\alpha = 0.995$, $d_{min} = 200$ frames and $\lambda(t)$ linearly decreasing from 4 to 2 along the test segment, brings an addition factor of 2 speed-up for the 5-20 sec. test conditions, with no difference on the development results. Figure 3 illustrates the evolution of the auto-adaptive pruning threshold on a test sample, in a case where an impostor provides a better likelihood than the true speaker at the beginning of the segment.

Overall, the cepstral features were computed at 0.1xRT. Target model adaptation was performed at 0.1xRT, and test identification at 0.08xRT down to 0.04xRT with pruning.

3.3 Developments Results

Developments were conducted on CHIL'05 Speaker Identification evaluation database, restricted to the microphone array matched condition, for the 30 seconds training condition and 1 to 30 seconds test segments. These are the most similar to CLEAR'06 conditions, despite the use of only 11 target speakers instead

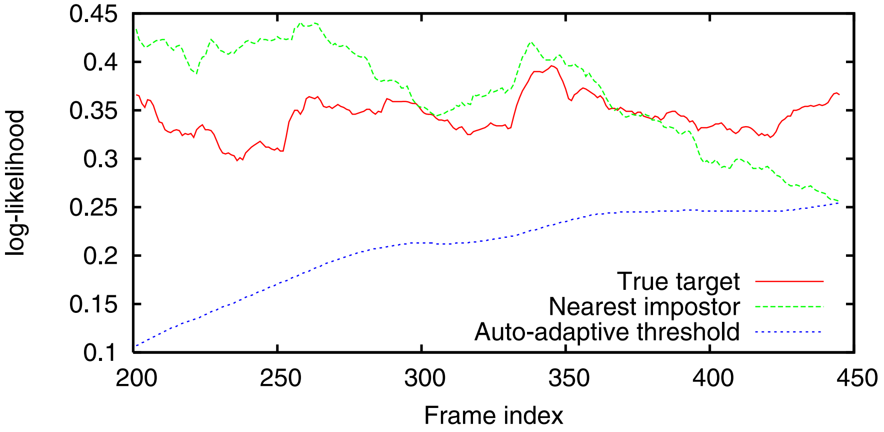

Fig. 3. Example of the evolution of the auto-adaptive pruning threshold during the recognition of a test segment

of 26. Results of LIMSI'05 system for CHIL'05 evaluation under these restricted conditions are reported Table 1. The system used an UBM with 2048 Gaussians trained on meeting data from various sources (ICSI, ISL, NIST) recorded using close-talking microphones, and cepstral mean and variance normalization was performed instead of feature warping [8]. The LIMSI'06 system provides a dramatic improvement for all segment durations, due mainly to better matched training data for the UBM. Contrastive experiments on feature normalization show that mean and variance normalization very significantly improve upon standard CMS, while feature warping is still slightly better. Other improvements to the system were mainly computation optimizations which do not show into the recognition scores.

3.4 CLEAR'06 Evaluation

Table 2 reports the LIMSI results for the CLEAR'06 evaluation. Note that for a few hundred trials, the precision of the identification error rates remain limited

Table 1. Identification error rates on the CHIL'05 Speaker Identification task restricted to microphone-array matched conditions, for the LIMSI'05 and the LIMSI'06 system associated with different feature normalizations

Test duration	*1 second*	*5 seconds*	*10 seconds*	*30 seconds*
# trials	1100	682	341	110
LIMSI'05	52.8	11.3	4.7	0.0
LIMSI'06 with CMS	33.4	5.6	1.8	0.9
LIMSI'06 with mean+variance	30.5	2.3	0.6	0.0
LIMSI'06 with feature warping	29.6	2.6	0.0	0.0

Table 2. LIMSI'06 system error rates for CLEAR'06 Acoustic Speaker Identification task

Test duration	*1 second*	*5 seconds*	*10 seconds*	*20 seconds*
# trials	613	411	289	178
Train A (15 seconds)	51.7	10.9	6.6	3.4
Train B (30 seconds)	38.8	5.8	2.1	0.0

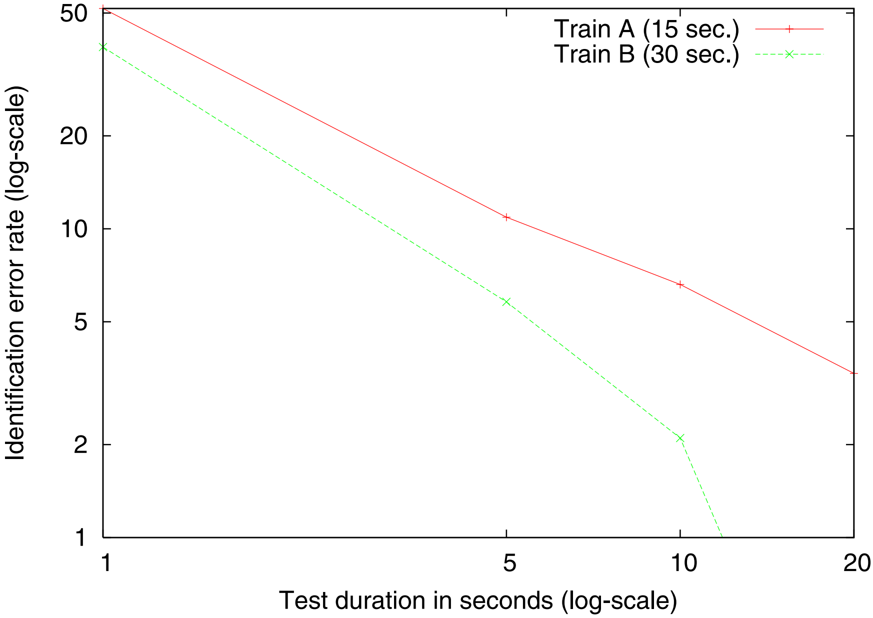

Fig. 4. LIMSI'06 system identification error rates by training and test duration for CLEAR'06 Acoustic Speaker Identification task

to ∼ 1%. The difference in speaker count does not allow a direct comparison with development results, but we can observe that the trends are similar. We observe especially high error rates on 1 sec. test segments. The effect of training and test durations are illustrated on a log-log scale in Figure 4.

4 Conclusions

The LIMSI CLEAR'06 system provides an over 50% relative reduction of the error rate compared to CHIL'05 Speaker Identification LIMSI results for a comparable configuration (matched array condition, 30 sec. training, 5 and 10 sec. test). Several optimizations were implemented and provided 10–20 acceleration factor in model training and speaker identification. The stochastic subsampling was shown to perform very efficiently compared to other existing approaches.

With the current system, no errors were measured for 30 sec. training and 20 sec. test segments; a larger test database would be necessary to increase the precision of the measure. However, identification rate of 1 second test segments remains poor compared to other results in the CLEAR'06 evaluation; our system would need specific tuning for very short segments.

Acknowledgments

Thanks are due to the CHIL partners for the seminar data, and in particular to ISL-UKA for making the audio beamforming available.

References

1. C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. of IEEE ICASSP*, May 2003.
2. G. Doddington, M. Przybocki, A. Martin, and D. Reynolds, "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225–254, 2000.
3. J.-L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2(2), pp. 291–298, April 1994.
4. J. McLaughlin, D. Reynolds, and T. Gleason "A Study of Computation Speed-UPS of the GMM-UBM Speaker Recognition System," in Proc. Eurospeech'99, pp. 1215–1218, Budapest, Sept. 1999.
5. D. Mostefa et al., "CLEAR Evaluation Plan v1.1," `http://isl.ira.uka.de/clear06/downloads/chil-clear-v1.1-2006-02-21.pdf`
6. J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Workshop on Speaker Recognition - Odyssey*, June 2001.
7. D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
8. X. Zhu, C-C. Leung, C. Barras, L. Lamel, and J-L. Gauvain, "Speech activity detection and speaker identification for CHIL," in *Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, Edinburgh, July 2005.