

The CLEAR'07 LIMSI System for Acoustic Speaker Identification in Seminars

Claude Barras, Xuan Zhu, Cheung-Chi Leung, Jean-Luc Gauvain, and Lori Lamel*

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

Abstract. This paper summarizes the LIMSI participation in the CLEAR 2007 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 working on cepstral coefficients, with MAP adaptation of a Universal Background Model (UBM). It builds on the LIMSI CLEAR'06 system with several modification: removal of feature normalization and frames filtering, and pooling of all speaker enrollment data for UBM training. The primary system uses a beamforming of all audio channel, while a single channel is selected for the contrastive system. This latter system performs the best and improves the baseline system by 50% relative for the 1 second and 5 seconds test conditions.

1 Introduction

Automatic person identification is a key feature of smart rooms, and in this context the European Integrated Project CHIL¹ has supported the CLEAR'06 and '07 evaluations, where audio, video and multi-modal person identification tasks were evaluated on CHIL seminars. Our work at LIMSI focuses on the acoustic modality. Similar to last year, the CLEAR'07 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 28 target speakers and test segment durations of 1, 5 10 and 20 seconds are considered².

This paper describes the LIMSI acoustic speaker identification system, evaluated in the CLEAR'07 benchmark. The system is a standard GMM-UBM system building on the LIMSI CLEAR'06 development [2]. In the next section, the LIMSI speaker recognition system is presented. 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/>

² <http://www.clear-evaluation.org/>

2 Speaker Recognition System

In this section, the LIMSI baseline speaker recognition system as proposed to CLEAR'06 evaluation and the new system developed for CLEAR'07 are described.

2.1 Baseline system

The speaker recognition system developed for the CLEAR'06 evaluation was taken as the baseline system for this year evaluation. It is organized as follows.

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. Ten percent of the frames with the lowest energy are filtered out, and short-term feature warping [4] is performed in order to map the cepstral feature distribution to a normal distribution.

A Gaussian mixture-model (GMM) with diagonal covariance matrices is used as a gender-independent Universal Background Model (UBM). This model with 256 Gaussians was trained on 90 min. of speech extracted from jun'04 and dev'06 CHIL data. For each target speaker, a speaker-specific GMM is trained by Maximum A Posteriori (MAP) adaptation [3] of the Gaussian means of the UBM. Target models are MAP-adapted using 3 iterations of the EM algorithm and a prior factor $\tau = 10$. 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 [5]. 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: $k^* = \operatorname{argmax}_k \log f(X|\lambda_k)$

Several optimizations addressing training and scoring computational requirements were implemented in the LIMSI CLEAR'06 system, in order to perform efficiently, in faster than real-time for realistic configurations. A stochastic frame subsampling was proposed for speeding up the UBM training using a large amount of training data. For the identification stage, top-Gaussian scoring was used, restricting the log-likelihood estimation to the 10 top scoring out of 256 components of the UBM for each frame and resulting to a 13 times speed up, and an auto-adaptive pruning was introduced, resulting in a further 2 times speed up for long duration segments.

2.2 System developed for CLEAR'07

For CLEAR'06 evaluation data, only the 4th out of the 64 channels of the MarkIII microphone array downsampled to 16kHz was used. Rather than picking a single channel, the ICSI beamforming software [1] was applied on the 64 channels for CLEAR'07 primary submission, with the 4th channel alone being used in the contrastive system. Neither feature normalization nor frame selection were used. Finally, the UBM was trained by pooling all speaker enrollment data

instead of using external data, which amounts to 7 minutes for the 15 seconds training condition and 14 minutes for the 30 seconds training condition. All other settings were kept unchanged.

3 Experiments

In this section the impact of the system changes on the CLEAR'06 evaluation and CLEAR'07 validation data are given. Results on the CLEAR'07 evaluation data are also provided for the primary and contrastive system.

3.1 CLEAR'06 evaluation

The results of LIMSI system in CLEAR'06 Acoustic Speaker Identification evaluation are reported in Table 1. The impact of two major changes in the system are given. Discarding feature normalization and UBM training by enrollment data pooling provide a dramatic improvement, an over 50% relative error reduction on the 1 and 5 seconds test condition.

Table 1. Identification error rates on the CLEAR'06 Speaker Identification task for the LIMSI'06 submitted system and for the modified system.

<i>Test duration</i>	<i>1 second</i>	<i>5 seconds</i>	<i>10 seconds</i>	<i>20 seconds</i>
LIMSI CLEAR'06 System				
Train A (15 s)	51.7	10.9	6.6	3.4
Train B (30 s)	38.8	5.8	2.1	0.0
+ no feature normalization				
Train A (15 s)	32.8	8.0	6.2	3.9
Train B (30 s)	20.1	3.4	2.4	1.1
+ enrollment data pooling for UBM				
Train A (15 s)	25.0	4.9	4.8	2.2
Train B (30 s)	16.2	1.9	0.7	0.0

3.2 Validation Experiments

Experiments were conducted using CLEAR'07 validation set in order to assess several settings of the system. Given the size of the validation set, only test durations of 1 and 5 seconds were considered as they provide respectively 560 and 112 samples; less than 100 samples were available for other test durations.

As was shown previously, the system is very sensitive to the feature normalization. Table 2 compares the identification error rate on the validation set for cepstral mean subtraction (CMS), mean and variance normalization

Table 2. Impact of various features normalizations (CMS, mean+variance, feature warping and raw features) on identification errors for beamformed and single channel audio, on CLEAR'07 validation data.

Normalization	Train/Test duration	Beamforming		4th channel	
		1 sec.	5 sec	1 sec.	5 sec
CMS	Train A (15 s)	38.8	6.2	46.4	11.6
	Train B (30 s)	28.7	3.6	38.0	4.5
mean+var	Train A (15 s)	39.6	2.7	49.8	13.4
	Train B (30 s)	30.2	2.7	37.7	3.6
warping	Train A (15 s)	39.6	2.7	48.6	9.8
	Train B (30 s)	28.6	0.9	39.6	4.5
raw	Train A (15 s)	17.9	2.7	21.1	3.6
	Train B (30 s)	14.1	1.8	15.5	1.8

(mean+var), feature warping and raw features. Avoiding any feature normalization is by far the best. This can be explained by a very limited channel variability per speaker in CHIL seminars. We can also note that beamformed audio is better in all configurations.

Table 3. Impact of UBM size and MAP prior weight on identification errors on CLEAR'07 validation data.

UBM size	MAP prior Train/Test duration	$\tau=8$		$\tau=10$		$\tau=12$	
		1 sec.	5 sec	1 sec.	5 sec	1 sec.	5 sec
128G	Train A (15 s)	17.7	6.2	18.2	6.2	19.1	6.2
	Train B (30 s)	14.8	0.9	14.6	0.9	14.6	0.9
256G	Train A (15 s)	17.9	2.7	17.9	2.7	17.7	2.7
	Train B (30 s)	14.3	1.8	14.1	1.8	14.5	0.9
512G	Train A (15 s)	20.7	4.5	20.7	3.6	20.7	3.6
	Train B (30 s)	14.3	1.8	14.1	1.8	14.3	1.8

Keeping raw features, we tested various UBM sizes (128, 256 and 512) and MAP adaptation weights (prior factor $\tau = 8, 10$ and 12) on the validation set with beamformed audio. As shown in Table 3, the baseline configuration with 256 Gaussians and $\tau=10$ remains a good compromise.

We finally checked the improvement provided by the frame selection. Table 4 gives the identification error rate with and without 10% low energy filtering on the validation set. Frames filtering does not seem to significantly help, except for 15 sec. training / 1 sec. test and was thus discarded from the final system.

Table 4. Identification errors on CLEAR’07 validation data with and without 10% low-energy frame filtering.

		<i>Beamforming</i>		<i>4th channel</i>	
<i>Filtering</i>	<i>Train/Test duration</i>	<i>1 sec.</i>	<i>5 sec</i>	<i>1 sec.</i>	<i>5 sec</i>
0%	Train A (15 s)	19.5	0.9	21.8	2.7
	Train B (30 s)	13.0	1.8	14.3	1.8
10%	Train A (15 s)	17.9	2.7	21.1	3.6
	Train B (30 s)	14.1	1.8	15.5	1.8

3.3 Evaluation Results

Table 5 reports the LIMSIS results for the CLEAR’07 evaluation for the primary and contrastive systems, along with CLEAR’06 results. We can observe that data beamforming, which was effective on validation set, did not work as expected in test condition. There may be some differences between validation and test data, and the settings of the beamforming were not optimized on the specific task configuration. There is less degradation for the contrastive system between the validation and test phases, between 25 and 30% relative. In CLEAR’06 evaluation, LIMSIS system showed rather high identification error rates on 1 sec. test segments, above 50% for 15 seconds training and near 40% for 30 seconds training. In ’07 contrastive system, these figures have been halved to 25% and 20% respectively. Both evaluation having a similar set of speakers (28 in CLEAR’07 vs. 26 in CLEAR’06), this allows a direct the comparison of the results. Figure 3.3 shows the improvement between LIMSIS ’06 and ’07 systems, as a function of the training and test durations in a log-log scale.

Table 5. LIMSIS error rates for CLEAR’06 and ’07 Acoustic Speaker Identification task.

<i>Test duration</i>	<i>1 second</i>	<i>5 seconds</i>	<i>10 seconds</i>	<i>20 seconds</i>
’06 Primary				
Train A (15 seconds)	51.7	10.9	6.6	3.4
Train B (30 seconds)	38.8	5.8	2.1	0.0
’07 Primary (beamforming)				
Train A (15 seconds)	37.6	9.2	6.2	2.7
Train B (30 seconds)	30.6	7.8	4.9	4.5
’07 Contrastive (4th channel)				
Train A (15 seconds)	25.0	5.1	3.1	1.8
Train B (30 seconds)	20.0	3.8	2.7	1.8

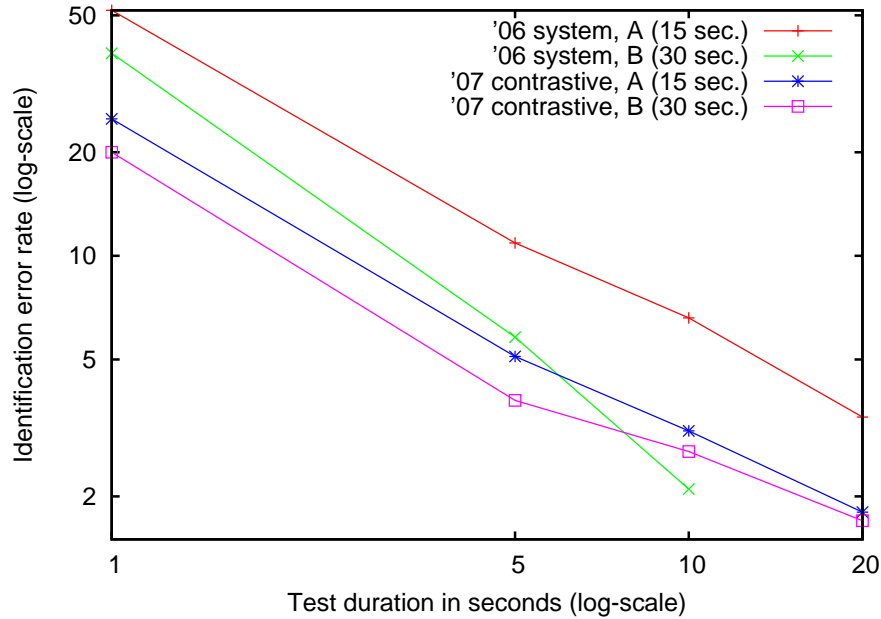


Fig. 1. Identification error rates by training and test duration for '06 and '07 contrastive LIMSI systems in CLEAR Acoustic Speaker Identification task.

4 Conclusions

The LIMSI CLEAR'07 contrastive system provides a 50% relative reduction of the error rate compared to previous year results in the 1 and 5 seconds test condition, resulting in 25% and 20% identification error rate for 15 and 30 second of training data, respectively.

These improvements are mainly due to modifications in the cepstral feature normalization step and in the UBM training. Feature warping usually improves speaker identification in telephone speech domain, and is also of interest for speaker diarization in broadcast news and meetings. However, discarding any feature normalization proved to be the most successful choice. This may be because a given speaker was generally recorded in a stable acoustic configuration along this evaluation. Training the UBM by pooling all enrollment data was chosen instead of using a large amount of available training data. This can only be considered in a closed-set speaker identification context, where the set of possible impostors is fully known in advance.

The primary system, taking advantage of a beamforming of all available 64 channels, does in fact degrade the performance in the test compared to the contrastive system where only a single channel is selected once for all. This observation is not consistent with the behavior of both systems on validation data, where beamforming always brought further improvement.

As a conclusion, the CLEAR'07 brought us a better insight into the speaker identification goals and constraints in the seminar meeting domain. This resulted in a dramatic improvement of the performances of our system in the short test conditions.

References

1. X. Anguera, C. Wooters, and J. Hernando, "Speaker Diarization for Multi-Party Meetings Using Acoustic Fusion", in *Automatic Speech Recognition and Understanding (IEEE, ASRU'05)*, San Juan, Puerto Rico, 2005.
2. C. Barras, X. Zhu, J-L. Gauvain, and L. Lamel, "The CLEAR'06 LIMSI Acoustic Speaker Identification System for CHIL Seminars", in R. Stiefelhagen & J. Garofolo editor, *Lecture Notes in Computer Science - CLEAR'06 Evaluation Campaign and Workshop - Classification of Events, Activities and Relationships*, Southampton, April 2006. vol. LNCS 4122, Springer Verlag, pp. 233-240, 2007.
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. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Workshop on Speaker Recognition - Odyssey*, June 2001.
5. D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.