Acoustic Speaker Identification: The LIMSI CLEAR'07 System

Claude Barras[1,2], Xuan Zhu[1,2], Cheung-Chi Leung[1], Jean-Luc Gauvain[1], and Lori Lamel[1,⋆]

[1] Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay cedex, France
[2] Univ Paris-Sud, F-91405, Orsay, France
{barras,xuan,ccleung,gauvain,lamel}@limsi.fr

Abstract. The CLEAR 2007 acoustic speaker identification task aims to identify speakers in CHIL seminars via the acoustic channel. The LIMSI system for this task consists of a standard Gaussian mixture model based system working on cepstral coefficients, with MAP adaptation of a Universal Background Model (UBM). It builds upon the LIMSI CLEAR'06 system with several modifications: 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 channels, 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 CHIL[1] 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 considered[2].

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 developments [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/

R. Stiefelhagen et al. (Eds.): CLEAR 2007 and RT 2007, LNCS 4625, pp. 233–239, 2008.

2 Speaker Recognition System

In this section, the LIMSI baseline speaker recognition system used in the 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 served as the baseline system for this year's 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: $k^* = \text{argmax}_k \log f(X|\lambda_k)$.

Several optimizations to reduce the training and scoring computational requirements were implemented in the LIMSI CLEAR'06 system in order to carry out identification 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 in a 13 times speed up, and an auto-adaptive pruning was introduced, resulting in a further factor of 2 speed up for long duration segments [2].

2.2 System Development for CLEAR'07

For CLEAR'06 evaluation data, only the 4th channel out of the 64 channels of the MarkIII microphone array was used. Rather than picking a single channel, the ICSI beamforming software [1] was applied to the 64 channels for CLEAR'07 primary submission, with the 4th channel alone being used in a contrastive system. For beamforming, the 1st channel was used as a reference for delay estimation, and other settings were kept identical to the default software configuration, with a delay estimation each 250ms on a 500ms window. In both cases the signal was

downsampled from 44kHz to 16kHz. 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 second training condition and 14 minutes for the 30 second training condition. All other settings were kept unchanged.

3 Experimental Results

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

3.1 Experiments with CLEAR'06 Evaluation Data

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 conditions.

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

Test duration	*1 second*	*5 seconds*	*10 seconds*	*20 seconds*
A: 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
B: A + 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
C: B + 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 Experiments with CLEAR'07 Validation Set

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; fewer 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 substraction (CMS), mean and variance normalization (mean+var),

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

		Beamforming		*4th channel*	
Normalization	*Train/Test duration*	*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

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

	MAP prior	τ=8		τ=10		τ=12	
UBM size	*Train/Test duration*	*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

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. It can also be noted that better results are obtained using beamformed audio data for all configurations.

Keeping raw features, tests were carried out varying the number of Gaussians in the UBM (128, 256 and 512) and the MAP adaptation weights (prior factor τ = 8, 10 and 12) on the validation set with the beamformed audio. As shown in Table 3, the baseline configuration with 256 Gaussians and τ=10 remains a good compromise.

In speaker identification, the GMM-UBM approach generally outperforms a direct training of the target models via maximum likelihood estimation (MLE). For contrastive purposes, identification performance on the validation set for MLE-trained models with a varying number of Gaussians are given in Table 4. The best results are obtained with 32 Gaussians for Train A (15 s) and with 64 Gaussians for Train B (30 s). These results are inferior to those obtained with the GMM-UBM configuration.

Table 4. Identification errors on the CLEAR'07 validation data using direct MLE trained models with a varying number of Gaussians

		Beamforming		*4th channel*	
GMM size	*Train/Test duration*	*1 sec.*	*5 sec.*	*1 sec.*	*5 sec.*
16G	Train A (15 s)	25.4	6.2	32.1	10.7
	Train B (30 s)	19.8	3.6	24.3	3.6
32G	Train A (15 s)	24.6	6.2	28.2	9.8
	Train B (30 s)	17.3	1.8	20.5	0.9
64G	Train A (15 s)	25.2	9.8	29.8	15.2
	Train B (30 s)	16.1	0.9	19.1	0.9
128G	Train A (15 s)	35.4	25.0	36.1	21.4
	Train B (30 s)	17.9	0.9	20.7	1.8

Table 5. Identification errors on the CLEAR'07 validation data with and without 10% low-energy frame filtering

		Beamforming		*4th channel*	
Filtering	*Train/Test duration*	*1 sec.*	*5 sec.*	*1 sec.*	*5 sec.*
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

The improvement provided by the frame selection was also assessed. Table 5 gives the identification error rate with and without 10% low energy filtering on the validation set. Frame filtering does not seem to significantly help, except for the 15 sec. training / 1 sec. test condition and was thus discarded from the final 2007 system.

3.3 CLEAR 2007 Evaluation Results

Table 6 reports the LIMSI results for the CLEAR'07 evaluation for the primary and contrastive systems, along with CLEAR'06 results, on the corresponding evaluation sets, expressed in terms of accuracy. It can be observed that data beamforming, which was effective on validation set, did not work as expected in the 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: given that a single speaker can be expected to be found in a segment, a single delay estimation on the whole segment between the reference and the other channels, as was done in [6], may have been a better choice.

There is less degradation for the contrastive system between the validation and test phases, between 25 and 30% relative. In CLEAR'06 evaluation, LIMSI system had rather low identification rates on 1 sec. test segments, below 50% for 15 seconds training and near 60% for 30 seconds training. In CLEAR'07 con-

Table 6. Accuracy rates for the LIMSI CLEAR'06 and '07 Acoustic Speaker Identification task on their respective evaluation data sets

Test duration	*1 second*	*5 seconds*	*10 seconds*	*20 seconds*
'06 Primary				
Train A (15 seconds)	48.3	89.1	93.4	96.6
Train B (30 seconds)	61.2	94.2	97.9	100.0
'07 Primary (beamforming)				
Train A (15 seconds)	62.4	90.8	93.8	97.3
Train B (30 seconds)	69.4	92.2	95.1	95.5
'07 Contrastive (4th channel)				
Train A (15 seconds)	75.0	94.9	96.9	98.2
Train B (30 seconds)	80.0	96.2	97.3	98.2

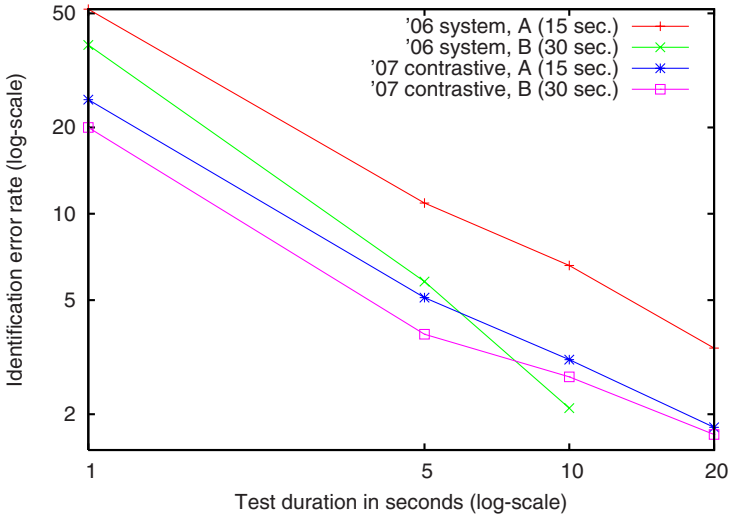

Fig. 1. Identification error rates by training and test duration for LIMSI '06 and '07 contrastive systems for CLEAR Acoustic Speaker Identification task

trastive system, these figures have been increased to 75% and 80% respectively. Both evaluations having a similar number of speakers (28 in CLEAR'07 vs. 26 in CLEAR'06), this allows a direct comparison of the results. Figure 1 shows the improvement between the LIMSI '06 and '07 systems, as a function of the training and test durations in a log-log scale.

4 Conclusions

LIMSI submitted two systems to the CLEAR'07 Acoustic Speaker Identification task. The contrastive system provides a 50% relative reduction of the error rate

compared to previous year results for the 1 and 5 seconds test conditions, resulting in 75% and 80% identification rate for 15 and 30 second training data, respectively.

This improvement is 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 for this evaluation. Training the UBM by pooling all enrollment data was chosen instead of using other 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. This configuration also outperformed a direct MLE training of the target models.

The primary system, taking advantage of a beamforming of all available 64 channels, performs substancially less well than the contrastive system where only a single channel is selected. This observation is different from the behavior of both systems observed on the validation data, where beamforming always outperformed a single channel. But the beamforming settings we used were not optimized for an array of distant microphones and for the specific evaluation conditions, so there is probably still room for system improvement in this area.

In conclusion, the CLEAR'07 has provided 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 for the short test conditions.

References

1. Anguera, X., Wooters, C., Hernando, J.: Speaker Diarization for Multi-Party Meetings Using Acoustic Fusion. In: Automatic Speech Recognition and Understanding (IEEE, ASRU 2005), San Juan, Puerto Rico (2005)
2. Barras, C., Zhu, X., Gauvain, J.-L., Lamel, L.: The CLEAR 2006 LIMSI Acoustic Speaker Identification System for CHIL Seminars. In: Stiefelhagen, R., Garofolo, J.S. (eds.) CLEAR 2006. LNCS, vol. 4122, pp. 233–240. Springer, Heidelberg (2007)
3. Gauvain, J.-L., Lee, C.H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing 2(2), 291–298 (1994)
4. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. In: Proc. ISCA Workshop on Speaker Recognition - Odyssey (June 2001)
5. Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19–41 (2000)
6. Luque, J., Hernando, J.: Robust Speaker Identification for Meetings: UPC CLEAR 2007 Meeting Room Evaluation System. LNCS, vol. 4625. Springer, Heidelberg (2008)