Robust Sound Source Identification for a Humanoid Robot

Alexej Swerdlow[1], Timo Machmer[1], Benjamin Kühn[1], Kristian Kroschel[1,2]

[1]Universität Karlsruhe (TH), Sonderforschungsbereich 588

[2]Fraunhofer-Institut für Informations- und Datenverarbeitung (IITB)

swerdlow@ira.uni-karlsruhe.de

Abstract: In this paper, investigations regarding the robust classification of acoustically observable sources (kitchen appliances and speakers) are presented. Thereby, two deciding factors are considered. On the one hand, the data acquisition should be exclusively done by the on-board sensors of the robot. On the other side, the entire information processing must be handled in real time due to the requirements given by the robot hardware. In so doing, there are positive and negative characteristics to think about. Since audio data are picked up in a small room like a kitchen with many sound sources, the background noise and reverberation of the signals of interest have to be taken into account. The redundancy of the signals picked up by the microphone array, which is installed on the robot head, can be employed to improve the classification accuracy.

The presented system is based on Gaussian Mixture Models in correspondence with the Mel Frequency Cepstral Coefficients as acoustic signal features. Furthermore, a Universal Background Model is used for the special case of speaker identification.

In this work, studies regarding the channel combination, the necessary length of the training phase, and the minimum data length for the evaluation are presented.

1 Introduction

There are a lot of areas, in which the robust identification of sound sources is required. One of them is the interaction between man and machine, which is given in scenarios, where a human interacts with a machine, for example a *humanoid robot*. Usually, the entire communication takes place via speech. In this case, the identification of speaking persons is of peculiar interest for the robot. But also in situations where no immediate contact between the user and the machine takes place, many other active sound sources can still exist in the robots proximity. A common example for this can be a kitchen, which contains different acoustically observable appliances. The robot ought to know its environment at any time to be able to find its way around. Especially, if handicapped or elderly people are involved, the humanoid robot has to guarantee the security of these people. Due to the reduced ability to hear, an elderly person might not register an acoustic event, so that the humanoid robot has to give a hint concerning it. Thus, the humanoid robot has to compensate the deficiency to hear of the person, who the robot takes care of.

Under controlled conditions, acoustic classification achieves very high performance. In a case of cross-validation, a correct recognition rate of 95 percent or even higher is usual. Also the annual **S**peaker **R**ecognition **E**valuation (SRE) of the **N**ational **I**nstitute of **S**tandards and **T**echnology (NIST) shows promising results [1]. However, the acoustic classification in a far-field scenario is still a great challenge. But exactly this situation occurs when a humanoid robot should recognize different sound sources in its proximity. In order to improve the recognition performance, different approaches were investigated for the last years. Some of them use advantages of distributed microphone arrays, as reported in [2]. Another method for

improving the recognition rate consists in the enhancement of the signal to noise ratio by using adaptive or fixed beamforming approaches, as described in [3], or by means of a directional microphone.

However, utilizing both techniques with the on-board sensors, which are placed on the head of the humanoid robot, crucial constraints have to be taken into account. For a start, audio data should be exclusively acquired by the on-board sensors of the robot and doing so there is no possibility to use distributed microphone arrays. Also the application of a beamforming approach seems not to be the ideal way. This conclusion results from two facts: at first, notable improvements are achieved by utilizing a large microphone array, for example consisting of 64 microphones, as reported in [3]. One has to accept, that the usage of such huge array is rather unrealistic for a head or even the body of a humanoid robot. But even if ignoring this fact, one more constraint has to be considered. At the latest, the computational costs of beamforming approaches do not allow their usage on the robot's hardware. The utilization of one or more directional microphone seems also not to be the best solution due to the required length of such kind of microphones.

On account of these considerations, the current paper presents a system for the robust acoustic classification of kitchen appliances and speakers, and tries to find an optimum parameter setup for its application in real environments, in association with a humanoid robot.

2 Context-independent sound source classification

Over the past years, the **M**el **F**requency **C**epstral **C**oefficients (MFCC) have proven to be the most appropriate parameters for speaker identification [4], which are also used as basic features for speech recognition. In our system we use the MFCCs for both speaker recognition and classification of kitchen appliances. The resulting versatility is the convenient advantage of this way of proceeding. In so doing, the sound signal is characterized by a 13-dimensional MFCC vector every 28ms. The first component of the feature vector reflects the energy of the analyzed speech segment and is not used for the classification. In order to exclude segments with no information, a sound activity detection based on normalized energy is utilized in the baseline system.

The individual speakers and various kitchen appliances can be distinguished on the basis of their specific acoustic features. Therefore an individual statistical model is required for each sound source. Over the past decades, **G**aussian **M**ixture **M**odels (GMMs) [5, 6] has become the method par excellence for the speaker identification task with text-independent speech data. We extended the GM modeling by adding the option for modeling of kitchen appliances as well. In order to determine the model parameters of the GMM for each sound source, a training phase is required. For this purpose, we drew on the **E**xpectation-**M**aximization (EM) algorithm [7, 6], which has proven to be the most efficient one. The parameters of GMMs are determined on the basis of the MFCC feature training vectors by the iterative application of the EM algorithm. The general GM modeling supports full covariance matrices. Contrary to that, we used diagonal covariance matrices only. For one thing this way of proceeding resulted in a higher computational efficiency, for another thing empirical investigations showed that diagonal-matrix GMMs normally outperform full matrix GMMs.

For the special case of speaker classification, the GM modeling can be replaced by a so called **U**niversal **B**ackground **M**odel (UBM) [8]. Using the UBM technique, the general class of speech is modeled by only one GMM for all different speakers. Instead of application of the EM algorithm for each speaker, individual speaker models are then derived from the UBM. Therefore, a form of Bayesian adaptation in combination with speaker-specific training vectors is used.

3 Experimental Setup

For the simplification of co-operation between humans and the robot, the entire communication is based on speech. In order to arrange the handling of the robot as flexible as possible, the microphones used for the acquisition of acoustic signals are fastened to the robot. The microphone array consists of six omni-directional condenser lavaliere microphones beyerdynamic MCE 60 (Figure 1). Two of them are placed on the positions of the human's ears, one on the forehead, one on the chin, and finally two further microphones are located on the back of the robot's head. The distance between the two ear microphones is 19 cm, between the front and back microphones 23 cm, between both front microphones 6 cm, and 4.5 cm between both microphones on the back, respectively.

Figure 1 - Microphone beyerdynamic MCE 60

In order to investigate the accuracy of the system for the robust acoustic classification, a sound source database was collected using all six microphones of the robot. We recorded speech data from ten speakers and five kitchen appliances with eight acoustically observable states in total. Thereby, we used a coffee grinder, a toaster, a bread cutter, a hand-held blender, and a household electric coffee machine with four acoustically observable states.

In order to consider the influence of different environments, real experiments were carried out in different test environments, and all recordings were done in two different typical office rooms. Each speaker was required to talk about topics of personal interest for about five minutes per room. That resulted in spontaneous free speech data of at least four minutes duration. Furthermore, all speakers were allowed to move around freely and were not forced to look towards the sensor array.

Analogously, all kitchen appliances were recorded in both rooms. The signal duration was at least two minutes for each appliance and each state, respectively. During the recording phase, all appliances were moved across the room.

4 Results

In this section, observational results and cognitions are presented. At this point, the attention should be paid again to the fact that our intention entailed in analysis of real sound data in real environments.

In order to evaluate the general system performance, measurements with recorded data from the same room (matched room conditions) were completed in the first step. For this purpose, a 10-fold-cross-validation was applied to the available speech data and a 3-fold-cross-validation to the kitchen appliances data, respectively. Due to the fact that the robot does not stay in the same room all the time, but navigates between different rooms, all measurements were repeated under mismatched room conditions, to wit: training in the first room and testing in the second room as well as vice versa. Subsequently, results from both rooms were averaged. In the following, for better comparison measurements are given for the cross-validation case (CV) as well as for mismatched room conditions (MM).

Furthermore, all results are presented for different signal acquisition durations, in particular for one, two, five, and ten seconds. That means that the classification result is available as soon as a sound data block of a specific length is acquired. The system performance is given by the classification accuracy, which is the percentage of correctly classified blocks over all blocks; the corresponding standard deviation (*std*) is given as well.

4.1 GMM size

At the first, basic system parameters had to be found. One of them is the number of Gaussians in the GM modeling. We evaluated GMMs with 8, 16, and 32 mixtures for both speakers and appliances. Additionally, GMMs with 64 mixtures were analyzed for the case of speaker classification.

Table 1 presents the average classification accuracy of the baseline system for both speakers and appliances, depending on the number of mixtures (results with the best parameter setup are highlighted in bold letters). It shows that accuracies under cross-validation conditions are much higher than under mismatched conditions, especially for the case of speaker classification. Thereby, the GMM training length was 120 seconds for speakers and 60 seconds for appliances, respectively.

As can be seen, the highest accuracy under mismatched conditions is achieved using 16 mixtures for speakers, and 8 mixtures for appliances. While 16 mixtures seems to be a good trade-off for both, restrictions concerning the computation efficiency motivated us to use different mixture numbers for speakers and appliances in all further measurements.

	GMM size / block length		8		16		32		64	
			avg	*std*	*avg*	*std*	*avg*	*std*	*avg*	*std*
Speakers (training: 120 s)	*1s*	*CV*	0.85	0.08	0.88	0.06	**0.89**	**0.06**	0.89	0.06
		MM	0.65	0.17	**0.68**	**0.17**	0.68	0.18	0.67	0.20
	2s	*CV*	0.93	0.05	0.94	0.04	**0.95**	**0.04**	0.95	0.04
		MM	0.73	0.18	**0.75**	**0.18**	0.75	0.20	0.74	0.22
	5s	*CV*	0.98	0.02	**0.98**	**0.02**	0.98	0.02	0.98	0.02
		MM	0.82	0.18	**0.85**	**0.17**	0.84	0.19	0.82	0.23
	10s	*CV*	0.99	0.01	**0.99**	**0.01**	0.99	0.01	0.99	0.01
		MM	0.86	0.19	**0.88**	**0.17**	0.86	0.19	0.84	0.24
Appliances (training: 60 s)	*1s*	*CV*	**0.95**	**0.05**	0.95	0.05	0.95	0.06	--	--
		MM	**0.92**	**0.06**	0.91	0.07	0.91	0.07	--	--
	2s	*CV*	0.96	0.04	**0.97**	**0.04**	0.96	0.05	--	--
		MM	**0.94**	**0.05**	0.94	0.06	0.93	0.06	--	--
	5s	*CV*	0.98	0.04	**0.98**	**0.03**	0.98	0.04	--	--
		MM	**0.97**	**0.03**	0.97	0.04	0.96	0.04	--	--
	10s	*CV*	0.98	0.04	**0.98**	**0.02**	0.98	0.04	--	--
		MM	**0.99**	**0.01**	0.98	0.02	0.98	0.03	--	--

Table 1 – Influence of the GMM size on the classification accuracy

4.2 Length of the training phase

As previously mentioned, for determining the model parameters of the GMM for each sound source, a training phase is required. That is why we evaluated the influence of the length of the training phase on the classification accuracy. Corresponding results are summarized in Table 2. As one can see, a much higher training length is required for speaker recognition, in comparison to the classification of kitchen appliances. While a training phase of not more than 60 seconds for kitchen appliances already leads to a rather high classification accuracy of 97% under mismatched conditions and a block length of five seconds, 120 seconds are

required to achieve an accuracy of 85% using the same block length for the case of speaker recognition.

	block length \ training length		15		30		60		90		120		180	
			avg	*std*	*avg*	*std*	*avg*	*std*	*avg*	*std*	*avg*	*std*	*avg*	*std*
Speakers (GMM size: 16)	*1s*	*CV*	0.73	0.12	0.80	0.10	0.85	0.08	0.86	0.07	**0.87**	**0.07**	--	--
		MM	0.51	0.18	0.59	0.18	0.64	0.18	0.66	0.18	0.67	0.17	**0.69**	**0.18**
	2s	*CV*	0.82	0.11	0.88	0.08	0.92	0.06	0.93	0.05	**0.94**	**0.04**	--	--
		MM	0.58	0.22	0.67	0.21	0.72	0.20	0.74	0.19	0.75	0.18	**0.77**	**0.18**
	5s	*CV*	0.89	0.09	0.95	0.06	0.97	0.04	0.98	0.03	**0.98**	**0.02**	--	--
		MM	0.65	0.25	0.75	0.24	0.81	0.20	0.83	0.20	0.85	0.17	**0.86**	**0.17**
	10s	*CV*	0.92	0.08	0.97	0.03	0.98	0.02	0.99	0.01	**0.99**	**0.01**	--	--
		MM	0.69	0.28	0.79	0.24	0.85	0.20	0.86	0.20	0.88	0.17	**0.89**	**0.16**
Appliances (GMM size: 8)	*1s*	*CV*	0.92	0.08	**0.95**	**0.06**	--	--	--	--	--	--	--	--
		MM	0.85	0.11	0.88	0.09	**0.92**	**0.06**	--	--	--	--	--	--
	2s	*CV*	0.94	0.07	**0.96**	**0.05**	--	--	--	--	--	--	--	--
		MM	0.87	0.10	0.90	0.08	**0.94**	**0.05**	--	--	--	--	--	--
	5s	*CV*	0.96	0.05	**0.98**	**0.03**	--	--	--	--	--	--	--	--
		MM	0.90	0.09	0.93	0.07	**0.97**	**0.03**	--	--	--	--	--	--
	10s	*CV*	0.96	0.06	**0.98**	**0.03**	--	--	--	--	--	--	--	--
		MM	0.92	0.08	0.95	0.06	**0.99**	**0.01**	--	--	--	--	--	--

Table 2 – Influence of the training length on the classification accuracy

4.3 Channel combination

As was mentioned above, the robot is equipped with a microphone array, which consists of six microphones. Our investigations concentrated on the redundancy of the signals picked up by the microphone array, with a view to improving the classification accuracy under mismatched conditions. For that purpose, we evaluated two different channel combination (CC) approaches.

In the first step, the channel combination was applied during the GMM training phase (CC training). That means that an individual model for each sound source was trained using sound data from all six microphones.

Subsequently, the channel combination can also be applied in another way (CC evaluation). Thereby, the total classification result is calculated by the logarithmic combination of the classification results over the desired block length, each of which is given by the evaluation of the trained GMM with sound data from different microphones. At this point, the attention should be paid to the fact that the corresponding GMM was trained with the sound data from one channel only (microphone on the forehead of the robot's head). The usage of separate GM models for each channel should result in moderate increasing of the classification accuracy.

Table 3 shows the system improvement by utilizing the two channel combination approaches under mismatched conditions. While channel combination does not yield much in case of appliances due to already good results without channel combination, the speaker classification task can mostly benefit from fusion of both channel combination approaches (CC both). For

example, a relative improvement of 9% could be achieved for sound data blocks of one second length.

	block length \ channel comb.		CC none		CC training		CC evaluation		CC both	
			avg	*std*	*avg*	*std*	*avg*	*std*	*avg*	*std*
Speakers	*1s*	*MM*	0.67	0.17	0.70	0.15	0.71	0.17	**0.73**	**0.16**
	2s	*MM*	0.75	0.18	0.78	0.16	0.79	0.17	**0.81**	**0.15**
	5s	*MM*	0.85	0.17	0.86	0.16	0.86	0.16	**0.87**	**0.16**
	10s	*MM*	0.88	0.17	0.90	0.15	0.89	0.15	**0.90**	**0.15**
Appliances	*1s*	*MM*	0.92	0.06	0.93	0.05	0.91	0.09	**0.94**	**0.06**
	2s	*MM*	0.94	0.05	**0.96**	**0.03**	0.93	0.09	0.96	0.05
	5s	*MM*	0.97	0.03	**0.98**	**0.02**	0.96	0.06	0.98	0.03
	10s	*MM*	**0.99**	**0.01**	0.99	0.01	0.97	0.04	0.98	0.02

Table 3 – Influence of the channel combination (CC) on the classification accuracy

4.4 UBM for speaker classification

As described in 2, a Universal Background Model (UBM) can be used for the restricted case of the speaker classification. In so doing, we trained a GMM with 64, 128, and 512 Gaussian mixtures, using approximately three hours of speech. The corresponding results with 60 seconds training for individual speaker models are summarized in Table 4. It shows that under mismatched conditions, the highest classification accuracy is achieved by using the UBM with 512 mixtures. In contrast, already the UBM with 64 mixtures seems to be sufficiently for the cross-validation case.

	block length \ UBM size		64		128		512	
			avg	*std*	*avg*	*std*	*avg*	*std*
Speakers (training: 60 s)	1s	*CV*	**0.84**	**0.11**	0.84	0.11	0.84	0.12
		MM	0.63	0.16	0.64	0.15	**0.66**	**0.14**
	2s	*CV*	**0.92**	**0.08**	0.92	0.08	0.91	0.09
		MM	0.72	0.18	0.73	0.16	**0.75**	**0.15**
	5s	*CV*	**0.97**	**0.05**	0.97	0.05	0.97	0.05
		MM	0.81	0.18	0.83	0.15	**0.85**	**0.14**
	10s	*CV*	**0.98**	**0.03**	0.98	0.03	0.98	0.03
		MM	0.85	0.20	0.87	0.15	**0.90**	**0.14**

Table 4 – UBM: influence of the GMM size on the classification accuracy

Figure 2 shows the influence of the UBM training length on the classification accuracy under mismatched conditions, in comparison to the common GM modeling. It can be clearly seen that the UBM approach results in significant increasing classification accuracy for short training phases. For 15, 60, and 120 seconds training length and a block length of five seconds, the relative improvement amounts 18.5%, 4.9%, and 2.4%, respectively.

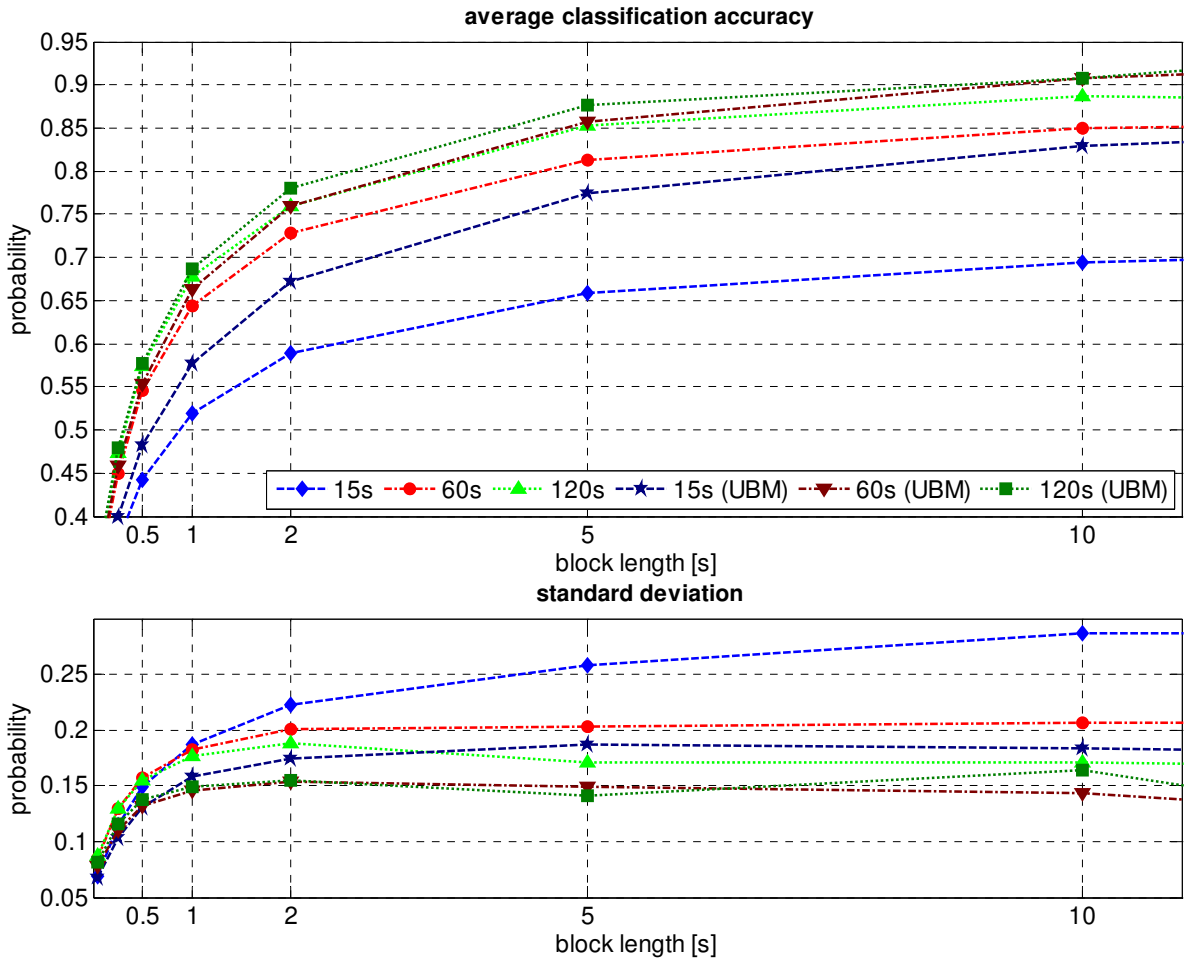

Figure 2 – In comparison: influence of different training lengths for common GMM and UBM

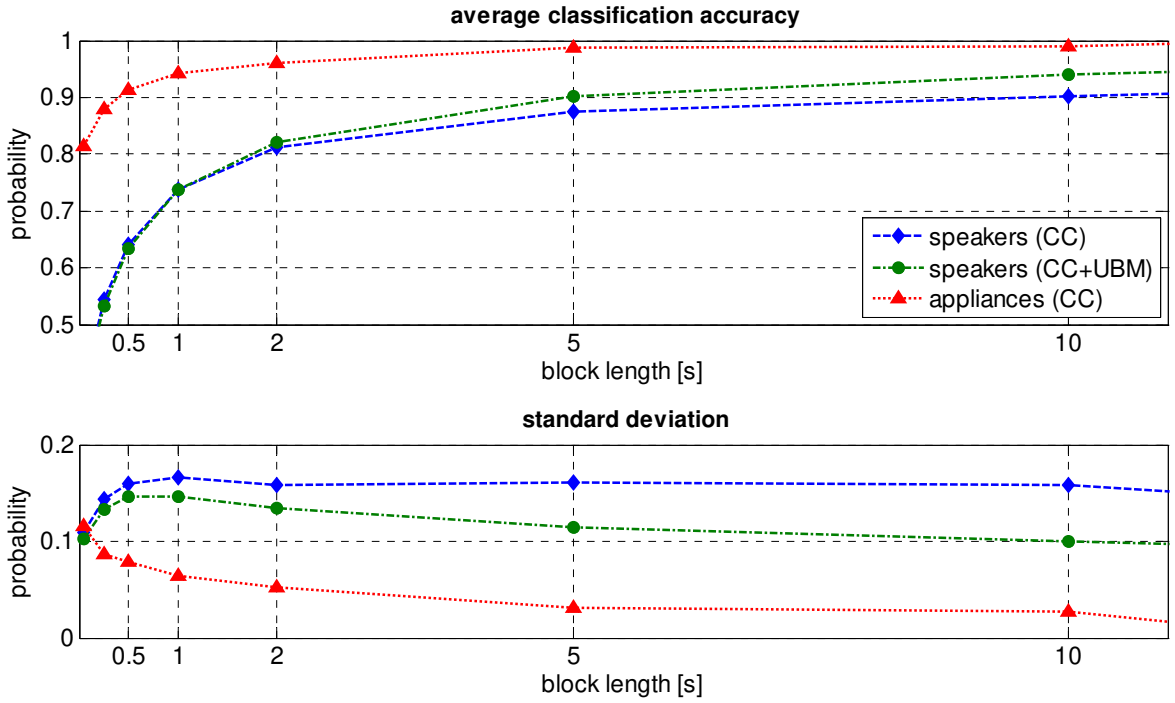

Figure 3 – Influence of the channel combination (CC) for common GMM and UBM

Finally, we evaluated the combination of the UBM technique with the channel combination approach. Figure 3 demonstrates the pleasant enhancement, which could be achieved by fusion of both methods under mismatched conditions. On average, a relative improvement of 4% was gained for blocks with at least two seconds of sound data. By way of comparison, the classification accuracy for appliances is given in the same figure.

5 Conclusion

In this paper, we presented our system for the robust classification of acoustically observable sources, both kitchen appliances and speakers. Real experiments were carried out in different test environments, and all recordings were done in two different typical office rooms. Subsequently, two main cases were differentiated: cross-validation within the same room and evaluation under mismatched conditions, to wit training with sound data from the first room and evaluation with sound data from the second room. This way of proceeding showed that in the cross-validation case classification accuracy is considerable high. Under mismatched conditions, the baseline system does not reach adequate results.

Our investigations pointed out the influence of the GMM size as well as the training length on the classification accuracy. In order to enhance the performance of the baseline system, two different channel combination techniques were proposed, whereby the combination of both approaches yielded. Furthermore, all results were given for different block lengths.

For the restricted case of the speaker classification, we showed that the UBM technique outperforms the common GM modeling, particularly with regard to short training phases. A fusion of the UBM approach with the channel combination technique resulted in the even higher classification accuracy.

The significance of different parameter setups for the classification of kitchen appliances and speakers was pointed out over the course of our evaluations.

Acknowledgement

This work has been supported by the German Science Foundation DFG within the Collaborative Research Center 588 “Humanoid Robots”.

Literature

[1] http://www.nist.gov/speech/tests/sre/

[2] Q. Jin, T. Schultz, and A. Waibel: "Far-field speaker recognition". Special Issue of the IEEE Transactions on Audio, Speech & Language on Speaker and Language Recognition. September 2007.

[3] C. Barras, X. Zhu, C.-C. Leung, J.-L. Gauvain, and L. Lamel: The CLEAR'07 LIMSI System for acoustic speaker identification in seminars. In R. Stiefelhagen, editor, Lecture Notes in Computer Science, Proc. CLEAR'07 Evaluation Campaign and Workshop - Classification of Events, Activities and Relationships, Baltimore, May 2007. Springer Verlag.

[4] D. O'Shaughnessy: Speech Communications - Human and Machine, IEEE Press, New York, 2000

[5] D. A. Reynolds, and R. C. Rose: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing, 3(1):73-83, January 1995

[6] D. Bechler, and K. Kroschel: Demonstrator zur automatischen textunabhängigen Sprechererkennung, 32. Deutsche Jahrestagung für Akustik (DAGA 2006), Braunschweig, March 2006

[7] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(1), 1977, 1-38

[8] D. A. Reynolds, T. F. Quatieri, and R. D. Dunn: Speaker verification using adapted Gaussian mixture models, Digital Signal Processing 10, 19–41, 2000