A FAST-MATCH APPROACH FOR ROBUST, FASTER THAN REAL-TIME SPEAKER DIARIZATION

Yan Huang[1,2], Oriol Vinyals[1], Gerald Friedland[1], Christian Müller[1], Nikki Mirghafori[1], Chuck Wooters[1]

[1]International Computer Science Institute, Berkeley
[2]Department of Computer Science, University of California, Berkeley
{yan,vinyals,fractor,cmueller,nikki,wooters}@icsi.berkeley.edu

ABSTRACT

During the past few years, speaker diarization has achieved satisfying accuracy in terms of speaker Diarization Error Rate (DER). The most successful approaches, based on agglomerative clustering, however, exhibit an inherent computational complexity which makes real-time processing, especially in combination with further processing steps, almost impossible. In this article we present a framework to speed up agglomerative clustering speaker diarization. The basic idea is to adopt a computationally cheap method to reduce the hypothesis space of the more expensive and accurate model selection via Bayesian Information Criterion (BIC). Two strategies based on the pitch-correlogram and the unscented-transform based approximation of KL-divergence are used independently as a fast-match approach to select the most likely clusters to merge. We performed the experiments using the existing ICSI speaker diarization system. The new system using KL-divergence fast-match strategy only performs 14% of total BIC comparisons needed in the baseline system, speeds up the system by 41% without affecting the speaker Diarization Error Rate (DER). The result is a robust and faster than real-time speaker diarization system.

***Index Terms*—** Speaker diarization, fast-match, pitch-correlogram, BIC, KL-divergence

1. INTRODUCTION

The goal of speaker diarization is to segment audio into speaker-homogeneous regions with the ultimate goal of answering the question "who spoke when?" [1]. Many state-of-the-art systems use a combination of agglomerative clustering with Bayesian Information Criterion (BIC) [2] and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs) [1][3]. These systems now obtain satisfactory accuracy in terms of speaker diarization error. However, the approach adopted in these systems exhibits inherent complexity due to the iterative cluster merging and sophisticated model selection procedure, which is often several times slower than real-time [4]. For most of the applications of speaker diarization, e.g. automatic speech recognition (ASR), large volume audio retrieval and multi-modal meeting event detection, faster than real-time performance is required.

In this paper, we present a fast speaker diarization approach by introducing a fast-match component to largely reduce the hypothesis space of the BIC-based model selection. The basic idea of fast-match is using a computationally cheap method to reduce the hypothesis space of the more expensive and accurate search, which has been widely used for word decoding in speech recognition [5]. Fast-match is essentially a search space tailoring technique.

Two fast-match strategies are explored in this work, each of which can be used separately. The first strategy uses the pitch-correlogram [6], as a type of prosodic feature, to capture speaker variances by looking at the pitch patterns. This technique has been successfully used for fast speaker recognition. In the second strategy we use KL-divergence, as a natural measurement of the difference between two probabilistic distributions. Although no closed-form expression exists for the KL-divergence between two GMMs, we utilize the accurate and efficient unscented-transform based approximation, which involves only evaluating likelihoods of Gaussian distributions at a few points and can achieve the approximation precision up to second order [7].

Based on these two strategies, we implemented two independent light-weight scoring schemes to measure how likely two clusters are to be merged before applying the more expensive model selection via BIC. Our proposed technique can reduce the hypothesis space by 86% and speed up the system by 41%. We achieve faster than real-time speaker diarization without affecting the speaker diarization error rate using an existing ICSI diarization system [3], which has obtained excellent results in past NIST evaluations.

The rest of this article is organized as follows: Section 2 introduces the framework of our fast-match approach for fast speaker diarization; Section 3 explains the pitch-correlogram and how it is used for fast-match in speaker diarization; Section 4 discusses the fast-match technique using the unscented-transform based approximation of KL-divergence; Section 5 shows the experiments and presents the results; Section 6 finally summarizes this article and points out future work.

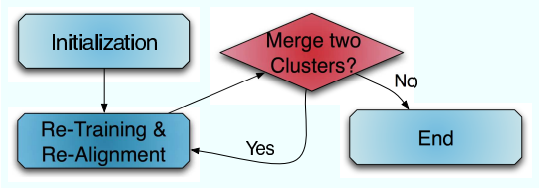

Fig. 1. Speaker diarization using agglomerative clustering, as explained in Section 2.

2. FAST-MATCH FRAMEWORK FOR FAST SPEAKER DIARIZATION

The agglomerative clustering approach used by many speaker diarization systems starts with a large number of initial clusters and proceeds by an iterative procedure of cluster merging, model re-training and re-alignment, as depicted in Figure 1. A more detailed description can be found in [7] [3].

In the cluster merging step, a merge score, which measures the goodness of model fitting using one merged model or two separate models based on Bayesian Information Criterion (BIC), is calculated between each two merge candidates. This measurement is then used to determine which two clusters should be merged or whether the merge should terminate. It terminates when no merging will improve the BIC score.

The computational load of such a system can be decomposed into three components: (1) find the best merge pair and merge; (2) model re-training and re-alignment; (3) other costs. After profiling the run-time distribution of an existing ICSI speaker diarization system [3], we find that the BIC score calculation takes 62% of the total run-time, as depicted in Table 1.

Component	Run-time
Find Best Merge Pair and Merge	62 %
Model Re-training/Re-alignment	28 %
Other	10 %
Total	100 %

Table 1. Run-time distribution of the ICSI speaker diarization system.

Analyzing how the best merge hypothesis is found, the reason for the high cost of the BIC score calculation can be identified. Let D_a and D_b represent the data belonging to cluster a and cluster b, which are modeled by θ_a and θ_b, respectively. D represents the data after merging a and b, i.e. $D = D_a \cup D_b$, which is parameterized by θ. The Merge Score (MS) is calculated as Eq. (1) [7]:

$$MS(\theta_a, \theta_b) = \log p(D|\theta) - (\log p(D_a|\theta_a) + \log p(D_b|\theta_b)) \quad (1)$$

For each merge hypothesis a and b, a new GMM (θ) needs to be trained. When the system is configured to use more initial clusters, which is preferable for better initial cluster purity, the computational load becomes prohibitive.

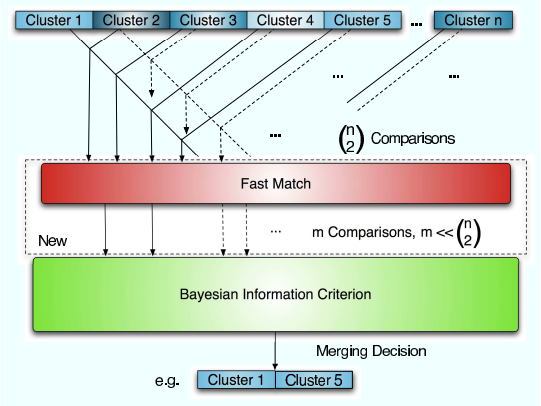

Fig. 2. Fast-match framework for fast speaker diarization, as explained in Section 2.

After identifying the BIC score calculation as the bottleneck for the whole system, we use a fast-match approach by introducing a new light-weight component to speed up the system as shown in Figure 2. This component is used as an intermediate step to determine the most likely merge hypotheses and only these merge hypotheses are passed to the more expensive BIC merge score calculation.

In the rest of this article, two strategies based on the pitch-correlogram and the unscented transform approximation of KL divergence between GMMs will be introduced. We would also like to point out that this framework is general and many other fast-match strategies can be used for speeding up a diarization system using agglomerative clustering with Bayesian Information Criterion. The main requirements are: the strategy should generate scores which roughly correlate with BIC scoring and it should be computationally efficient.

3. PITCH-CORRELOGRAM FAST-MATCH APPROACH

Speech is normally thought of as a physical process consisting of a sound source (i.e. the vocal chords) and a channel, which includes the vocal tract, the tongue, lips and etc. This is also known as the speech production model [8]. Pitch analysis tries to capture the fundamental frequency of the sound source. A pitch-correlogram [6] can be used to capture the variations of pitch dynamics among different speakers when sufficient data is available.

3.1. Pitch-correlogram

The pitch-correlogram is used to capture pitch dynamics by looking at the statistics of pitch patterns at frame level distance [6]. Specifically, a pitch-correlogram (H) is the joint distribution of quantized pitch bands explored at certain frame

level distances (k). When k is set to 1, the pitch-correlogram is a 2-dimensional table $H = [h_{ij}]$, which basically collects the bigram statistics of the quantized pitch of neighboring frames as shown in Eq. (2):

$$h_{i,j} = \frac{\#(i,j)}{n-1}, \quad \sum_{i,j=1}^{M} h_{i,j} = 1, \tag{2}$$

where $\#(i,j)$ counts the occurrences of the i-th band followed by the j-th band among all neighboring frames and n is the total number of frames.

Pitch varies by $2\% - 10\%$ in successive voiced frames, which implies that the transition of the pitch bands between neighboring voiced frames is lazy in the sense that they are dominated by small band change. H turns out to be a very sparse matrix with a lot of zero values on off-diagonal entries. Human pitch roughly ranges between 50-500 Hz (or in logarithm pitch domain $\log 50 - \log 500$). We use 110 bins to linearly quantize this pitch region (or logarithm of this pitch region).

The pitch-correlogram generation only involves counting bigrams of pitch bands and can be calculated efficiently. The rest of this section will discuss three different kinds of distance used to measure the distance between two pitch-correlograms.

3.2. Distance measure

Since the pitch-correlogram is essentially a histogram of quantized pitch bigrams, as mentioned before, the quantitative measure of their dissimilarity can be calculated using bin-by-bin histogram dissimilarity measurements as summarized in [9].

Three different kinds of bin-by-bin dissimilarity measurements are adopted, i.e. Minkowsky-form distance (d_{L_r}) , histogram intersection distance ($\hat{d_{\bigcap}}$) and Jeffrey divergence (d_J), with details in [9]. We empirically found the Jeffrey divergence (d_J) performs best as will be shown in Section 5.4. Suppose H_i and H_j are the pitch-correlogram for cluster i and j, $h_i^{(k)}$ is the k-th bin of the histogram H_i. Jeffrey divergence (d_J) is calculated as:

$$d_J(H_i, H_j) = \sum_k (h_i^{(k)} \log \frac{h_i^{(k)}}{m^{(k)}} + h_j^{(k)} \log \frac{h_j^{(k)}}{m^{(k)}}), \tag{3}$$

where $m^{(k)} = \frac{h_i^{(k)} + h_j^{(k)}}{2}$.

3.3. Limitation of the pitch-correlogram

The pitch-correlogram exhibits certain speaker discriminability and captures speaker variances. It is well suited for our fast-match framework for fast speaker diarization or speaker grouping for hierarchical speaker recognition. However, it is not recommended to use it as a standalone speaker indexing feature in applications where a large pool of speaker impostors exists as described in [6].

4. KL-DIVERGENCE FAST-MATCH APPROACH

Our second proposed strategy, instead of utilizing new features (eg. prosodic), as in our first approach, uses the original low-level cepstral features with Gaussian mixture modeling. It measures how likely it is that two models are to be merged by asking the question, "how different are these two distributions" via KL-divergence between two GMMs approximated using the highly accurate and efficient unscented transform.

4.1. KL-divergence

The KL-divergence of two distributions is defined as

$$KL(f(x)||g(x)) = \int f(x) \log \frac{f(x)}{g(x)} dx \tag{4}$$

and the symmetric version is

$$\hat{KL}(f(x)||g(x)) = KL(f(x)||g(x)) + KL(g(x)||f(x)). \tag{5}$$

When $f(x)$ and $g(x)$ are Gaussian mixture distributions, there is no closed-form expression. To solve the integration in Eq. (4), sampling methods, such as Monte-Carlo simulation, can be used. The Monte-Carlo method generates a sequence of sampling points, simulating the distribution, and approximates the integration by performing summation over this simulated sequence. However, this is computationally expensive and does not suit our need for efficiency.

4.2. Efficient and highly accurate approximation using the unscented-transform

The unscented transformation is a method for calculating statistics of a random variable undergoing a nonlinear transform [10], which looks similar to Monte-Carlo sampling since it also generates sampling points, but is fundamentally different in that the samples are generated in a completely deterministic fashion.

Suppose X is a d-dimensional random variable with expectation μ_X and x_s are the sigma points chosen from X according to Eq. (6), whose sample mean equals μ_X and sample covariance equals $\sum_X$:

$$x_s(k) = \mu_X \pm \sqrt{d[\sum_X]_k}, \quad k = 1, \cdots, d, \tag{6}$$

where $[\sum_X]_k$ is the k-th column of the covariance matrix $\sum_X$.

If Y is a new random variable generated by applying a nonlinear transformation Q to X, i.e. $Y = Q(X)$, then μ_Y can be approximated by the sample mean of y_s, which are the nonlinear transform of the sigma point x_s, i.e. $y_s = Q(x_s)$, and this approximation is precise up to second order [10]. The approximation of KL-divergence between Gaussian mixture models based on the unscented-transform has been used in speaker recognition applications [11].

Applying the unscented transform, $\int f(x)\log g(x)dx$ can be approximated using Eq. (7) [11], which is sufficient for illustrating the approximation for Eq. (5):

$$\int f(x)\log g(x)dx = \frac{1}{2d}\sum_{i=1}^{n}\alpha_i\sum_{k=1}^{2d}\log g(x_{i,k}). \quad (7)$$

Since a diagonal covariance matrix is used (as in our speaker diarization system), i.e. $\sum_i = diag(\sigma_{i,1}^2,\cdots,\sigma_{i,d}^2)$, the sigma points are simply:

$$\mathbf{x}_{i,k} \quad = \mu_i \pm \sqrt{d}\sigma_{i,k}e_k, \quad i = 1,\cdots,n, k = 1,\cdots,d, \quad (8)$$

where e_k is a d-dimensional indicator vector and has all zero components except one only at the k-th entry. This approximation is precise up to second order from the theorem of the unscented-transform [10]. Since it only involves evaluating the likelihood of Gaussian at a few sigma points, it is computationally efficient.

5. EXPERIMENTS AND RESULTS

5.1. Baseline system

The baseline system used in our experiments is an existing ICSI Meeting Evaluation development system [12], which has three components: Feature Extraction (FE), Speech Activity Detection (SAD) and Speaker Clustering (SC). The total system run-time is the sum of the speaker clustering run-time (T_{SC}) and the run-time of feature extraction and speech activity detection (T_{FE+SAD}). Since the new fast-match system does not change the first two components, T_{FE+SAD} is constant for both systems.

The data set contains 12 meetings and a total of 2.3 hours (T) of audio data to be diarized. This is a set we put together as the development set for NIST Meeting Evaluation 2006[4], which will be referred to as DEV06 data set in the rest of this paper. The diarization engine only processes the speech frames output by SAD with total duration (T_{SP}).

To better illustrate the effect of our fast-match approach, we use two real-time measurements for system run-time: (1) $xRT_1 = T_{SC}/T_{SP}$; (2) $xRT_2 = (T_{SC}+T_{FE+SAD})/T$. The first one measures the real time factor of the speaker diarization engine for speaker clustering, the second one measures the real time factor of the whole diarization system. All experiments were performed on an Intel Xeon 2.8 GHz machine with 512 KB of cache and 3 GB of RAM, which was exclusively reserved for these experiments. The operating system used was Linux Red Hat Enterprise 4.

The baseline system has a DER of 11.74% and performs 4951 BIC comparisons in total with BIC score based heuristic pruning described in [12]. Table 2 summarizes this baseline system:

$DER = 11.74, \#BIC = 4951$					
T	T_{SP}	T_{SC}	T_{FE+SAD}	xRT_1	xRT_2
8150s	6972s	11173s	640s	1.60	1.45

Table 2. Summary of the baseline speaker diarization system: the total meeting time (T) and the total speech time (T_{SP}); the run-time for speaker clustering (T_{SC}) and the run-time for FE and SAD (T_{FE+SAD}); the real-time factor for speaker clustering (xRT_1) and for speaker diarization (xRT_2).

5.2. Matching rate (r)

The theoretical number of merge hypotheses at each iteration is $n_i(n_i-1)/2$, where n_i is the number of clusters at the i-th iteration and $n_i = n_{i-1} - 1$. The matching rate (r) is used to control the shrinking rate of the hypothesis space after fast-match, which is defined as the ratio of the total number of comparisons needed after and before fast-matching.

Since the hypothesis space decreases after each iteration, we find it is necessary to dynamically boost the matching rate so that we start with a more constrained matching and become more "relaxed" as the size of hypothesis space decreases.

Two types of boosting for matching rate (r) are used: the first one always matches the top m candidates at each iteration, which implicitly increases r since the size of hypothesis space decreases along iterations; the second approach explicitly linearly decreases the matching rate r as in Eq. (9):

$$r_i = r_0 + i * \frac{(1-r_0)}{M}, i = 0,\cdots,M-1, \quad (9)$$

where M is the initial number of clusters and r_0 is the initial matching rate.

5.3. Average cross ranking percentage (R) and matching chance (MC)

Besides using the NIST DER to measure the performance of the diarization system, we also introduce the average cross ranking percentage (R) and the Matching Chance (MC) to measure the goodness of the new scoring and the effectiveness of fast-matching.

The average cross ranking percentage (R) measures the average ranking percentage of the top m-ranked pairs hypothesized by BIC scoring using the new scoring (F) as shown in Eq. (10)

$$R_{F,BIC} = \frac{\sum_{i=1}^{m} rank_F([rank_{BIC}]_i)}{\#cmp}, \quad (10)$$

where F denotes the new scoring scheme, $rank_F$ is the ranking using F, $[rank_{BIC}]_i$ is the index of the top i-th hypothesis by BIC ranking and $\#cmp$ is the total number of BIC-comparison without fast-matching.

The Matching Chance (MC) is the probability that the fast-match will not mistakenly exclude the best BIC proposed hypothesis among all iterations and all meetings.

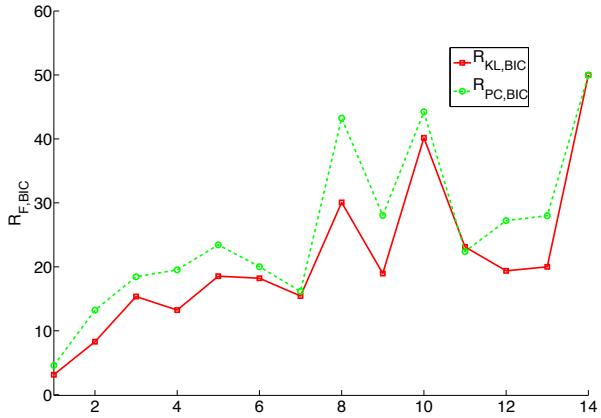

Fig. 3. The average cross ranking percentage of pitch-correlogram ($R_{PC,BIC}$) and KL-divergence ($R_{PC,BIC}$), as defined in Section 5.3.

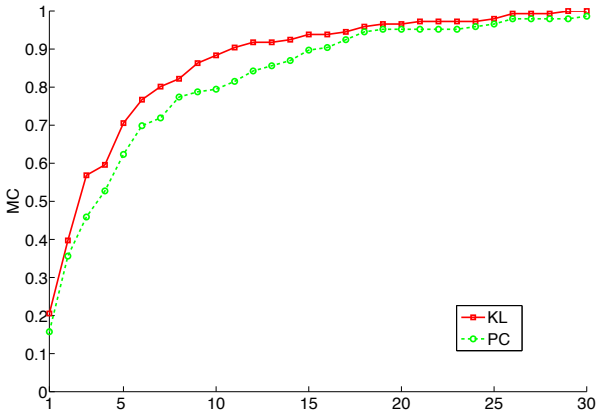

Fig. 4. The matching chance (MC) of pitch-correlogram and KL-divergence, as defined in Section 5.3.

5.4. Results

In order to choose the distance measurement for pitch-correlograms, we experimented with three different distance measurements: Minkowsky-form distance (d_{L_r}), symmetric histogram intersection distance ($\hat{d}_{\bigcap}$) and Jeffrey divergence (d_J), as introduced in Section 3.2. The linear boosting of matching rate as shown in Eq. (9) is used with the initial matching rate set to 10%. Table 3 shows the diarization results of fast-match using pitch-correlogram. Jeffrey divergence (d_J) performs better than the other distance measurements in terms of DER and is used as the distance measurement for pitch-correlograms in the rest of this paper.

Distance	d_{L_r}	$\hat{d}_{\bigcap}$	d_{L_J}
DER (%)	12.64	12.18	11.49

Table 3. Diarization results of pitch-correlogram fast-match versus different distance measurements (d_{L_r}, $\hat{d}_{\bigcap}$ and d_{L_J}), as explained in Section 3.2.

The average cross ranking percentage of the pitch-correlogram ($R_{PC,BIC}$) and KL-divergence ($R_{KL,BIC}$) versus BIC is shown in Figure 3, which depicts the average ranking percentage of the top 1 BIC decisions in the new scoring versus iterations. As can be observed from this figure, in order to keep the top 1 BIC decision in the match space, the matching rate (r) needs to be increased along iterations as discussed previously. We can also see that the KL-divergence scoring has a better average cross ranking percentage than the pitch-correlogram scoring.

Figure 4 shows the matching chance (MC) of the pitch-correlogram and KL-divergence versus the number of the merge candidates (m) kept after fast-match. When $m = 1$, it is equivalent to discarding BIC scoring completely and only using it in making merge decisions, and the matching chance (MC) is only 20%. But as m increases to 15, the matching chance increases to 80∼90%. We would also like to point out that since it is not completely clear that the BIC decision is optimal in terms of DER, the MC-curve does not necessarily exactly correlate with DER.

The results of the diarization speed-up and DER are shown in Table 4. The pitch-correlogram approach, with the starting match rate set to 10% with linear boosting along iterations as in Eq. (9), speeds up the system by 22% without degrading the DER. The best result is achieved using the KL-divergence approach, which matches only the top 5 candidates at each iteration. It performs 715 BIC comparisons (14% of 4951 comparisons performed in the baseline system), speeds up the system by 41% and achieves faster than real-time speaker diarization without affecting the speaker diarization error rate. We can also see that in the same setup for the pitch-correlogram approach, even though the system is sped up by 40%, there is a 0.79% degredation in DER. This partially verifies our statement on the limitation of the pitch-correlogram in Section 3.3. Since in the KL-divergence fast-match approach, the system performance is not affected by keeping only the top 5 merge candidates at each iteration, we move one step further to see what if we only keep the top 1 candidates. In this case the BIC score is not used to decide which two clusters should be merged, but only used to decide

	DER	$\#BIC$	T_{SC}(s)	xRT_1	xRT_2	SU
Baseline	11.74%	4951	11173	1.60	1.45	NA
$PC_{r=0.1}$	11.49%	2035	8727	1.25	1.15	22%
PC_{Top5}	12.53%	697	6667	0.96	0.90	40%
$KL_{r=0.1}$	12.52%	2060	8347	1.20	1.10	25%
KL_{Top5}	11.58%	715	6570	0.88	0.94	41%

Table 4. Results of pitch-correlogram and KL-divergence fast-match (SU is the speedup of T_{SC} over the baseline).

when merging should stop. The diarization result is 19.56%. This shows that BIC, as a robust model selection technique, still makes better merge decisions than simple distance measurements.

6. CONCLUSION AND FUTURE WORK

We presented a fast-match approach for agglomerative speaker diarization. The fast-match component reduces the hypothesis space of the expensive model selection using BIC. Two fast-match strategies, which are used separately, are adopted to reduce the number of BIC comparisons, which was identified as the bottleneck of the speaker diarization system based on agglomerative clustering.

The first strategy, as a feature level strategy, uses the inexpensive pitch-correlogram to capture speaker variation by looking at the pitch patterns in order to pre-select highly likely merge candidates. The second strategy, as a model level strategy, uses the efficient and highly accurate unscented-transform based approximation of KL-divergence to measure the distance of two models. These two fast-match strategies, as two independent approaches, are used to select the most likely merge candidates.

The pitch-correlogram fast-match approach performs less than half of the total BIC comparisons of the baseline and speeds up the system by 22% without degrading the DER. The best performance is achieved by using the KL-divergence fast-match, which reduces the number of BIC comparisons by 86% and speeds up the existing ICSI speaker diarization system by 41% without affecting the accuracy. The result is a robust and faster than real-time speaker diarization system, which makes its integration in upstream applications more practical and efficient.

Since the pitch-correglogram only carries the speech source information, while the KL-divergence approach contains the speech channel difference via MFCC feature, it would be interesting to combine these two strategies for further speedup via more aggressive fast-match. However, an effective way to do this is still to be researched.

7. ACKNOWLEDGMENTS

The authors would like to thank Marijn Huijbregts and Nelson Morgan for helpful discussion in this work.
This work was (partly) funded by DTO VACE. Oriol Vinyals was supported by the European Union 6th FWP IST Integrated Project AMIDA (Augmented Multiparty Interaction with Distant Access). Gerald Friedland and Christian Müller were supported by a fellowship within the postdoc program of the German Academic Exchange Service (DAAD).

8. REFERENCES

[1] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings of the IEEE International Conference on Audio and Speech Signal Processing*, 2005.

[2] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings of DARPA speech recognition workshop*, 1998.

[3] B. Peskin X. Anguera, C. Wooters and M. Aguilo, "Robust speaker segmentation for meetings: The icsi-sri spring 2005 diarization system," in *Proceeding of the NIST MLMI Meeting Recognition Workshop, Edinburgh*, 2005.

[4] M. Huijbregts and C. Wooters, "The blame game: Performance analysis of speaker diarization system components," in *Proceedings of the Interspeech*, 2007.

[5] D. Kanevsky L. Bahl, P.S. Gopalakrishnan and D. Nahamoo, "Matrix fast match: a fast method for identifying a short list of candidate words for decoding," in *Proceedings of the IEEE International Conference on Audio and Speech Signal Processing*, 1989.

[6] N. Jhanwar and A. K. Raina, "Pitch correlogram clustering for fast speaker identification," *EURASIP Journal on Applied Signal Processing*, pp. 2640–2649, 2004.

[7] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proceedings of the IEEE Automatic Speech Recognition Underdtanding Workshop*, 2003.

[8] G Fant, "Acoustic theory of speech production," Mouton De Gruyter, 1970.

[9] C. Tomasi Y. Rubner and L. J. Guibas., "A metric for distributions with applications to image databases," in *Proceedings of the IEEE International Conference on Computer Vision*, 1998.

[10] S. Julier and J. Uhlmann, "A general method for approximating nonlinear transformations of probability distributions," Technical Report, RRG, Dept. of Engineering Science, University of Oxford, 1996.

[11] H. Aronowitz J. Goldberger, "A distance measure between gmms based on the unscented transform and its application to speaker recognition," in *Proceedings of Interspeech*, 2005.

[12] B. Peskin C. Wooters, J. Fung and X. Anguera, "Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system," in *Proceedings of NIST Rich Transcription Workshop*, 2004.