SPEAKER DIARIZATION OF MEETINGS BASED ON LARGE TDOA FEATURE VECTORS

Deepu Vijayasenan[1], Fabio Valente[2]

[1] Universität des Saarlandes, Saarbrücken, Germany
[2] Idiap Research Institute, 1920, Martigny, Switzerland
deepu.vijayasenan@lsv.uni-saarland.de,fabio.valente@idiap.ch

ABSTRACT

This paper investigates the use of large TDOA feature vectors together with acoustic information in speaker diarization of meetings. TDOAs are obtained by considering all possible microphones pairs and this approach is compared with conventional TDOA features extracted w.r.t. a reference channel. The study is carried using two systems, the first based on Gaussian Mixture Modeling and the second based on the Information Bottleneck approach. Results on NIST RT06/RT07/RT09 evaluation datasets show a large speaker error reduction of 30% relative going from 14.3% to 10.8% for the first and from 12.3% to 8.2% for the second whenever the feature weighting is properly handled. Furthermore results reveal that the IB system is more robust to different number of microphones even when all pairs large TDOA vectors are used thus outperforming the HMM/GMM by 25% relative (8.2% error compared to 10.8%).

Index Terms— Speaker diarization, Time Delay Of Arrival features, Meetings Recordings, Model combination.

1. INTRODUCTION

Speaker diarization is an unsupervised learning task with the objective of finding *"who spoke when"* in a given audio recording. In recent years, diarization has been applied to meeting recordings acquired using Multiple Distant Microphones (MDM) and several methods have been proposed to effectively use the redundancy coming from the MDM audio. Beamforming techniques have been investigated where the various audio sources are merged to produce a single high-quality audio stream [1]. The beamforming algorithm [1] selects a reference channel based on the average cross-correlation and performs a Delay-and-Sum combination. The Time Delay of Arrivals (TDOA) are estimated with respect to the reference channel. Besides beamforming, the TDOA features also carry information about the location of the current speaker and they have been used as complementary features to conventional MFCC [2]. The combination happens at model level, weighting the log-likelihoods of independent GMM models estimated on each feature stream.

However, the TDOA feature statistics and quality is influenced by several factors like the number of microphones in the array (variable for each recording environment), the acoustic of the room, reverberation/noise and the relative position of speakers respect to the array (see [3] for analysis). The choice of estimating delays according to a single reference channel, chosen as the one that has the highest average cross-correlation over the entire recording, may be locally suboptimal since the TDOA is the result of the different speakers placements with respect to the microphones.

In [4], authors proposed to compute TDOA between all microphone pairs resulting into a large vector and to select only the five pairs that have the highest peak-to-peak difference thus most representative of speakers position. After that, a with-in pairs and an across pairs quantization step is applied in order to find nine clusters used as initialization into a conventional diarization system. This approach produced state-of-the-art performances during the Rich Transcription 2009 evaluation. Instead of selecting the best pairs, those large TDOA vectors have also been reduced to one or two components by means of Principal Component Analysis or Discriminant Analysis [5] with the drawback of noisy covariance matrix estimations. Motivated by the performances of those large TDOA feature vectors obtained by considering all microphone pairs in providing a good system initialization [4, 6], this paper investigates their use as complementary features to MFCC for diarization. In contrary to [3, 4, 5, 6], no selection, dimension reduction nor TDOA based initialization is performed and the use of the entire all-pairs delay vector is investigated. The main challenge comes from the increased dimensionality and this work will focus on how the new vector dimension affects the combination with MFCC features. The study is carried using two state-of-the-art diarization systems, the first based on HMM/GMM modeling and the second based on the Information Bottleneck (IB) principle - a non parametric clustering framework.

The remainder of the paper is organized as follows: section 2 describes the delay feature estimation, sections 3 and 4 describe the state-of-the-art diarization systems used in this study while section 5 presents experiments and analysis of the two systems. The paper is concluded in section 6.

2. DELAY FEATURE ESTIMATION

TDOA features are estimated using the generalized cross correlation phase transform (GCC-PHAT) [1]. All time delays are calculated with respect to a reference channel. This channel is chosen based on the signal to noise ratio or depending on the average cross correlation of the channel with other channels. After choosing a reference channel, signal in each channel is windowed using a $500ms$ window. Given two windowed signals $x_i(n)$ and $x_j(n)$, the GCC-PHAT is defined as :

$$G_{PHAT}(f) = \{X_i(f)X_j^*(f)\}/\{|X_i(f)||X_j(f)|\} \quad (1)$$

where $X_i(f)$ and $X_j(f)$ are the Fourier transforms of the two signals. The TDOA for these channels is estimated as

$$d_{PHAT}(i,j) = \arg\max_d R_{PHAT}(d) \quad (2)$$

where $R_{PHAT}(d)$ is the inverse Fourier transform of $G_{PHAT}(f)$.

Thus given M microphones in the array, the dimension of the TDOA feature vector is $M-1$; in the following we will refer to those as *reference channel* TDOA features. On the other hand, whenever all possible microphones pairs are considered, i.e., Eq. (2) is computed for all pairs (i,j), the dimension of the TDOA vector becomes $\frac{1}{2}M(M-1)$ thus much larger then the previous one; in the following we will refer to those as *all pairs* TDOA features.

3. HMM/GMM DIARIZATION

This section briefly describes a conventional speaker diarization system based on HMM/GMM models in which each speaker is represented by an HMM state with GMM emission probability [7]. Let us designate the emission probability distribution b_{c_k} of cluster c_k with $\log b_{c_k}(s_t) = \log \sum_r w^r_{c_k} \mathcal{N}(s_t, \mu^r_{c_k}, \Sigma^r_{c_k})$ where s_t is the input feature, $\mathcal{N}(.)$ is the Gaussian pdf and $w^r_{c_k}$, $\mu^r_{c_k}$, $\Sigma^r_{c_k}$ are the weights, means and covariance matrices (diagonal) corresponding to r^{th} mixture Gaussian of cluster c_k. The diarization starts with a uniform linear segmentation of the input into a large number of clusters (speakers). Successively, at each step, a cluster pair is merged based on a distance measure like the BIC or its modified version [7]. The merging stops when all the BIC values are less than zero. After each merge, a Viterbi realignment of speaker boundaries is performed with the estimated speaker models. Whenever multiple feature streams $\{s^i_t\}$, e.g., MFCC $\{s^{mfcc}_t\}$ and TDOA $\{s^{tdoa}_t\}$ (extracted at the same rate) are available, the system can be extended by considering a separate GMM model for each stream (see [2]). Let $b^i_c(s^i_t)$ be the GMM model of cluster c corresponding to the feature stream s^i_t. A separate GMM emission distribution $b^i_{c_k}(.)$ is estimated for each feature stream. A combined log likelihood is then computed for each cluster c_k as:

$$\log L_{c_k}(s_t) = W_{mfcc} \log \left[b^{mfcc}_{c_k}(s^{mfcc}_t) \right] + W_{tdoa} \log \left[b^{tdoa}_{c_k}(s^{tdoa}_t) \right] \quad (3)$$

where, W_i corresponds to the weight of each feature stream ($W_{mfcc} + W_{tdoa} = 1$). This combined likelihood $\log L_{c_k}(.)$ replaces the log likelihood terms $\log b_{c_k}(.)$ during clustering and realignment (see [2, 8] for details). It can be noticed that the log-likelihoods in Eq. 3 are dependent on the dimension of the feature vectors s^{mfcc} and s^{tdoa} thus increasing the dimension of the TDOA vector from $M-1$ to $M(M-1)/2$ will increase the magnitude of the second term in Eq. 3 ($\log \left[b^{tdoa}_{c_k}(s^{tdoa}_t) \right]$).

4. INFORMATION BOTTLENECK DIARIZATION

This section briefly summarizes the Information Bottleneck speaker diarization system that operates in a normalized space of relevance variables proposed in [9]. The Information Bottleneck is a distributional clustering technique introduced in [10]. Consider a set of input variables X. The Information Bottleneck principle depends on a relevance variables' set Y that carries important information about the problem. According to IB principle, any clustering C should be compact with respect to the input representation (minimum $I(X,C)$) and preserve as much mutual information as possible about relevance variables Y (maximum $I(C,Y)$). This corresponds to the maximization of:

$$\mathcal{F} = I(C,Y) - \frac{1}{\beta} I(X,C) \quad (4)$$

where β is a Lagrange multiplier. The IB criterion is optimized w.r.t. the stochastic mapping $p(c|x)$ using iterative optimization techniques. The agglomerative Information Bottleneck (aIB) clustering is a greedy way of optimizing the IB objective function [10]. The algorithm is initialized with each input element $x \in X$ as a separate cluster. At each step, two clusters are merged such that the reduction in mutual information w.r.t relevance variables is minimum. It can be proved that the loss in mutual information in merging any two clusters c_1 and c_2 is given in terms of a Jensen-Shannon divergence that can directly be computed from the distribution $p(y|x)$ in closed form. The number of clusters is determined by using a threshold on the Normalized Mutual Information given by $\frac{I(C,Y)}{I(X,Y)}$.

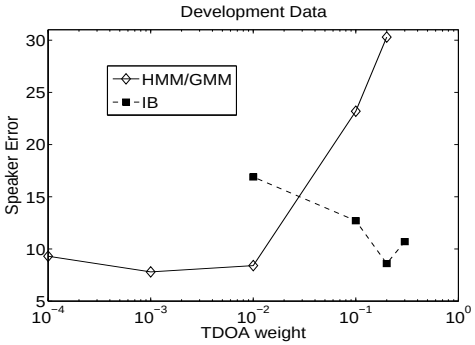

Fig. 1. Speaker error function of all-pairs TDOA weights on development data in case of HMM/GMM and IB Speaker Diarization. Similar graph for the reference channel TDOA can be found in [2, 11].

In order to apply this method to speaker diarization, the set of relevance variables $Y = \{y_n\}$ is defined as the components of a background GMM ($\mathcal{M}$) trained on the entire audio recording [9]. The input to the clustering algorithm is uniformly segmented speech segments x_t. The posterior probability $p(y_n|x_t)$ is computed using Bayes' rule. The speech segments with the smallest distance (the Jensen-Shannon divergence) are then iteratively merged until the model selection criterion is satisfied.

Whenever multiple features are available, the combination is performed in the space of relevance variables Y [11]. Separate GMMs with the same number of components are trained for each feature stream. The individual components are kept aligned. i.e, the same component of two different GMMs are estimated using the features with same time indices. In other words, there is a one-to-one correspondence between the GMM components. Let $\{\mathcal{M}^{mfcc}, \mathcal{M}^{tdoa}\}$ be the background model for the MFCC and TDOA feature vectors. The combined distribution $p(y|x)$ for each segment x^{mfcc} and x^{tdoa} is then estimated as:

$$p(y|x) = W_{mfcc} p(y|x^{mfcc}, \mathcal{M}_{mfcc}) + W_{tdoa} p(y|x^{tdoa}, \mathcal{M}_{tdoa}) \quad (5)$$

This corresponds to averaging the different $p(y|x^i, \mathcal{M}_i)$ obtained with GMMs trained on different feature streams. After clustering, the speaker boundaries are realigned. Instead of using HMM/GMMs, the realignment is performed in the space of relevance variables $p(y|x)$ using a Kullback-Leibler divergence based HMM system described in [11].

The entire diarization algorithm including clustering, feature combination and realignment depends only on the relevance variable distribution $p(y|x)$. Most importantly, we can notice that combination (see Eq.5) is performed with probabilities rather then log-likelihoods as in HMM/GMM diarization, thus being less affected by the dimension of TDOA vector.

5. EXPERIMENTS

The experiments are conducted on 24 meeting recordings from six different meeting rooms (CMU,EDI,NIST,IDI,TNO,VT) corresponding to data collected for the NIST RT06/RT07/RT09 evaluations [12]. The meetings identifier as well as the number of microphones associated with each meeting are reported in Table 1. At first, multiple channels are beamformed using the *BeamformIt* toolkit [13]. MFCC features are then extracted from the beamformed output (details about the front-end are available in [8]). Delays are obtained both with respect to a reference channel as well as using all possible microphone pairs. The system performance is evaluated using Diarization Error Rate (DER) that is the sum of speech/non-

Table 1. Meeting number, identifier and associated number of microphone for each recording.

ID	Meet.	#Mic	ID	Meet.	#Mic	ID	Meet.	#Mic
1	CMU_20050912-0900	2	*9*	EDI_20071128-1000	16	*17*	NIST_20080201-1405	7
2	CMU_20050914-0900	2	*10*	EDI_20071128-1500	16	*18*	NIST_20080227-1501	7
3	CMU_20061115-1030	3	*11*	IDI_20090128-1600	16	*19*	NIST_20080307-0955	7
4	CMU_20061115-1530	3	*12*	IDI_20090129-1000	16	*20*	TNO_20041103-1130	10
5	EDI_20050216-1051	16	*13*	NIST_20051024-0930	8	*21*	VT_20050408-1500	4
6	EDI_20050218-0900	16	*14*	NIST_20051102-1323	8	*22*	VT_20050425-1000	7
7	EDI_20061113-1500	16	*15*	NIST_20051104-1515	7	*23*	VT_20050623-1400	4
8	EDI_20061114-1500	16	*16*	NIST_20060216-1347	7	*24*	VT_20051027-1400	4

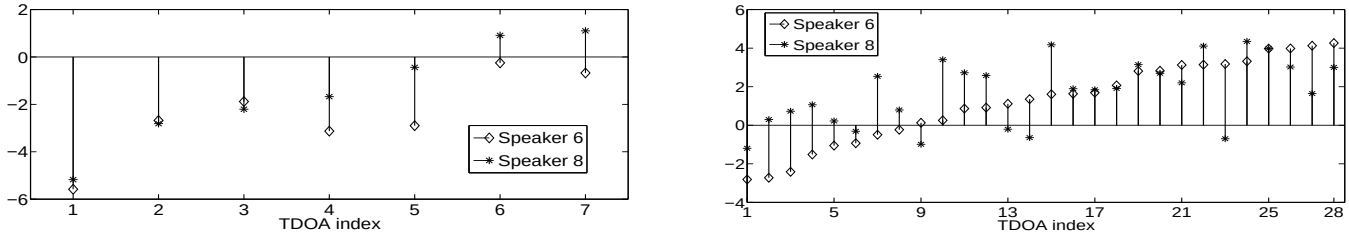

Fig. 2. Average TDOA values in case of speaker 6 and 8 in $IDI\_20090129-1000$ meeting (NIST reference notation). Left plot represents TDOA w.r.t. a reference channel while right plot represents features w.r.t. all possible pairs in case of a single microphone array (8 microphones).

speech segmentation and speaker errors. Since we use the same speech/non-speech segmentation across all the experiments, only speaker error is reported for the purpose of comparison. The combination weights W_{mfcc} and W_{tdoa} (see Eq. 5 and Eq. 3) are estimated from a development dataset composed of recordings across 6 meetings rooms as in the test data set. The weights are selected as those that minimize the speaker error on the development data set. Whenever conventional reference channel delay features are used, the typical HMM/GMM (W_{mfcc}, W_{tdoa}) weighting is $(0.9, 0.1)$. This conventional system is considered as baseline in this work. As described in section 3, the model log-likelihood is proportional to the feature vector dimension thus moving from a delay vector of dimension $M-1$ to a vector of dimension $M(M-1)/2$, will increase the delay feature log-likelihood. Keeping the $(0.9, 0.1)$ weighting produces a speaker error above 20% (see Figure 1) on development data. In order to compensate for this effect, weights (W_{mfcc}, W_{tdoa}) are optimized on a logarithmic scale. Figure 1 (solid line) reports the performance of the HMM/GMM on development data when the weighting is optimized on a a logarithmic scale: it can be noticed that, when the TDOA feature vector moves from $M-1$ to $M(M-1)/2$, the optimal weights move from $(0.9, 0.1)$ to $(0.999, 0.001)$. In other words, the effect of increase in dimensionality can be compensated by tuning the feature weights (W_{mfcc}, W_{tdoa}) in a logarithmic scale.

In case of conventional TDOA features, the typical (W_{mfcc}, W_{tdoa}) weights for IB diarization are $(0.7, 0.3)$. Figure 1 (dashed line) also reports the performance of the IB on development data in case of all-pairs TDOA: it can be noticed that when the TDOA feature vector moves from $M-1$ to $M(M-1)/2$, the optimal weights move from $(0.7, 0.3)$ to $(0.8, 0.2)$. The increased dimensionality only marginally affects the weighting as the combination is done using probabilities (see Eq. 5) in the space of relevance variables. Table 2 summarizes the weightings in case of conventional TDOA feature as well as all-pairs TDOAs while Table 3 reports the speaker error obtained using such weightings on the evaluation dataset. It can be noticed that, both in case of HMM/GMM and IB, the error is reduced by more than 30% relative achieving speaker errors equal to 10.8% and 8.2% respectively. Interestingly, optimizing the weights on a logarithmic scale, make the HMM/GMM system benefit of those large feature vectors without the need of selecting the best pairs as in [3] nor reducing the dimensionality as in [5].

Table 2. Weighting for MFCC and TDOA feature vectors in case of all-pairs TDOA and reference channel TDOA obtained minimizing the speaker error on development data set.

	aIB	HMM/GMM	TDOA dim.
Ref. Channel TDOA	(0.7,0.3)	(0.9,0.1)	M-1
All Pairs TDOA	(0.8,0.2)	(0.999,0.001)	$M(M-1)/2$

Table 3. Speaker Error obtained by HMM/GMM and IB diarization on evaluation data set; TDOA features are computed respect to a reference channel and as all possible TDOA pairs.

	aIB	HMM/GMM
Ref. Channel TDOA	12.3	14.3
All Pairs TDOA	8.2 (+33%)	10.8 (+32%)

For analysis purposes, let us plot the average TDOA values in case of meeting $IDI\_20090129$-1000 (only one array out of the two available, i.e., 8 microphones) for two speakers (speaker 6 and 8 according to NIST reference files) that are merged together by both systems into a single cluster. Figure 2 (left figure) plots the TDOA feature values estimated w.r.t. a reference channel and the values of the all-pairs TDOA features (right figure). In the first case only two features exhibit a sign change, while other five features have same sign and almost comparative values. In the second case, 13 features out of 28 have different signs making the diarization distinguish better in between the two speakers that are confused by the reference channel TDOA thus suggesting that, depending on the speakers location, the choice of a single reference channel can be suboptimal.

As previously pointed in [11], the IB system appears more robust to variation of weights across different meeting recordings. This robustness holds also in case of all-pairs TDOA feature vectors. Figures 3 and 4 plot the meeting-wise error for the 24 recordings in Table 1 that compose the evaluation data set in case of HMM/GMM and IB systems. From figure 3, it can be noticed that in case of HMM/GMM, improvements happen on recordings with larger number of microphones (7 or more) - while the new features/weights produce some degradation in case of recordings performed with less than 4 microphones (meetings ID 1,2,3,4,21,23,24). On the other hand, Figure 4 shows that in case of IB diarization, improvements are verified on recordings with larger number of microphones (7 or

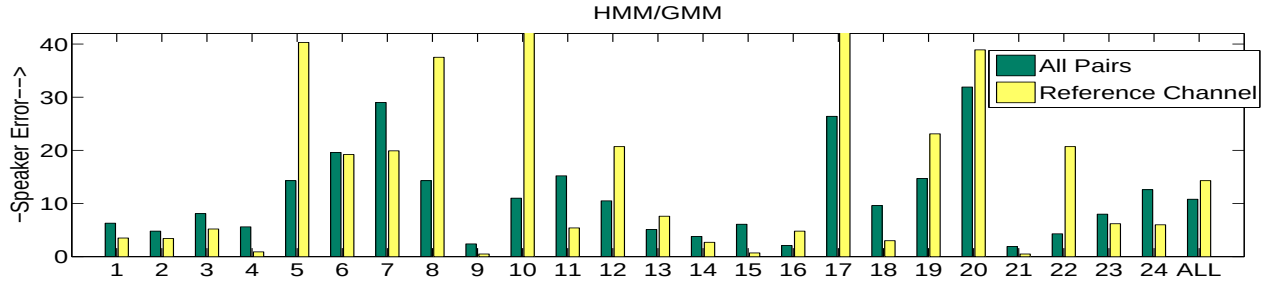

Fig. 3. Speaker Error obtained by HMM/GMM diarization the RT06/RT07/RT09 data set whenever TDOA features are computed respect to a reference channel (yellow bars) and whenever all possible TDOA pairs are computed (green bars)

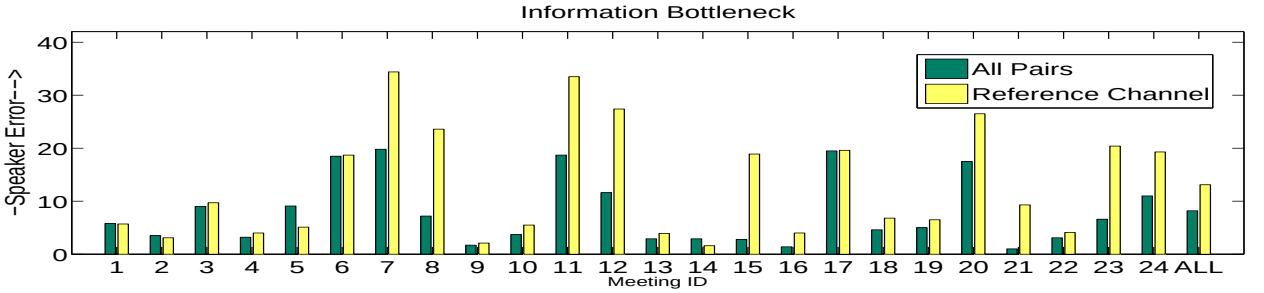

Fig. 4. Speaker Error obtained by IB diarization the RT06/RT07/RT09 data set whenever TDOA features are computed respect to a reference channel (yellow bars) and whenever all possible TDOA pairs are computed (green bars)

more) with not much degradation on others as feature streams are combined using normalized quantities (see Eq. 3) instead of log-likelihoods (see Eq. 5) which are dependent on the feature dimension.

6. CONCLUSION AND DISCUSSIONS

Many state-of-the art diarization systems combine acoustic information with TDOA features computed with respect to a reference channel from the michrophone array. Previous works have shown that this choice can be suboptimal depending on the position of the speakers respect to the michrophone, e.g. [3], thus proposing the use of TDOA values computed for all possible microphones pairs. Issues related to the increased dimensionality of the vector have been addressed selecting the most performant pairs [4] or reducing the dimensionality of the vector [5] before using them for diarization purposes. This work investigates how those large TDOA vector can be directly used in diarization systems and studies their combination with acoustic information using two systems: a parametric HMM/GMM system and the IB system. Experiments on 24 meetings from the RT06/RT07/RT09 NIST RT evaluations reveal that all-pairs TDOA features become effective in HMM/GMM modeling only when the combination weights are optimized on a *logarithmic* scale in order to compensate for the increased dimensionality. In this case the speaker error is reduced by $+32\%$ relative (from 14.3% to 10.8%) w.r.t. conventional delay features and, interestingly, no need for pair selection [4] nor dimensionality reduction [5] is needed.

Whenever IB diarization is performed, the increased dimensionality marginally affects the optimal weighting and, also in this case the speaker error is reduced by $+32\%$ relative (from 12.3% to 8.2%). This effect is due to the fact the IB system combines the information in a normalized space of relevance variables. Furthermore, also in case of large TDOA vectors, the IB system outperforms the HMM/GMM being more robust to dimensionality variations achieving a speaker error of 8.2% compared to 10.8%. In summary, whenever weighting issues are properly handled, diarizing with delays obtained using all possible michrophone pairs can reduce the speaker error by $+30\%$ relative compared to conventional delays computed respect to a reference channel, the improvement being larger with increasing number of microphones.[1].

7. REFERENCES

[1] Anguera X., Wooters C., and Hernando J., "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 7, 2007.

[2] J.M. Pardo , X. Anguera, C. Wooters, "Speaker Diarization For Multiple-Distant-Microphone Meetings Using Several Sources of Information," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1189, 2007.

[3] Koh E.C.W., Sun H., New T.L., Nguyen T.H., Bin M., Li H., and Rahardja S., "Speaker diarization using direction of arrival estimate and acoustic feature information: The i2r-ntu submission for the nist rt 2007 evaluation," in *Lecture Notes of Computer Science Vol. 4625, Multimodal Technologies for Perception of Humans*, 2008.

[4] Sun H., Nwe T.L., Bin M., and Li H., "Speaker diarization for meeting room audio," in *Proc. of Interspeech*, 2009.

[5] Evans N., Fredouille C., and Bonastre J.F., "Speaker diarization using unsupervised discriminant analysis of inter-channel delay features.," in *Proceedings of ICASSP*, 2009.

[6] Sun H., Ma B., Khine S.Z.Z., and Li H., "Speaker Diarization System for RT07 and RT09 Meeting Room Audio," *Proceedings of ICASSP*, 2010.

[7] Jitendra Ajmera, *Robust Audio Segmentation*, Ph.D. thesis, Ecole Polytechnique Federale de Lausanne (EPFL), 2004.

[8] Xavier Anguera, *Robust Speaker Diarization for Meetings*, Ph.D. thesis, Universitat Politecnica de Catalunya, 2006.

[9] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 7, pp. 1382 – 1393, 2009.

[10] N. Slonim, N. Friedman, and N. Tishby, "Agglomerative information bottleneck," in *Proceedings of Advances in Neural Information Processing Systems*. MIT Press, 1999, pp. 617–623.

[11] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard, "An information theoretic combination of mfcc and tdoa features for speaker diarization of meetings data," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 2, 2011.

[12] "http://www.nist.gov/speech/tests/rt/rt2006/spring/," .

[13] X. Anguera, "Beamformit, the fast and robust acoustic beamformer," in *http://www.icsi.berkeley.edu/xanguera/BeamformIt*, 2006.

[1] Authors would like to thank Dr. X. Anguera for his help with the Beamformit toolkit. This work was funded by the EU Seventh Framework Programme ([FP7/2007-2013] undergrant agreement n [213850] in the SCALE project, the EU NoE SSPNet and the SNF NCCR IM2.