SAPA-SCALE Conference
Portland, OR, USA
September 7-8, 2012

Dimensionality Reduction of Large TDOA Vectors for Speaker Diarization

Deepu Vijayasenan[1], *Fabio Valente*[2]

[1]Universität des Saarlandes, 66123 Saarbrücken, Germany
[2]Idiap Research Institue, 1920 Martigny, Switzerland
dvijayasenan@lsv.uni-saarland.de, fvalente@idiap.ch

Abstract

In this work, we investigate a dimensionality reduction scheme to use Time Delay of Arrival(TDOA) features across all microphones in a traditional HMM/GMM system. The subspace dimension is selected based on dimension of the TDOA vectors in an ideal recording, i.e., without environmental distortion or interference. Experiments in a dataset used in NIST Meeting Diarization evaluation reveal that the dimensionality reduction to a considerably lower dimension improve the diarization error by 3.7%(30% relative). While the proposed scheme has the advantage that it does not require any development set tuning to select the dimension as proposed by previous methods, it retains competitive performance (5% better than tuning the results).

Index Terms: Speaker diarization, Time Delay of Arrival, Dimensionality reduction.

1. Introduction

Speaker Diarization refers to the unsupervised learning task of "*who spoke when*". Given an audio recording, speaker diarization involves determining the number of speakers as well as the speech segments corresponding to each speaker. A common approach to diarization is to iteratively merge an over-segmented set of speech segments in a bottom up fashion [1] using the mel frequency cepstrum as features. The audio is often captured using Multiple Distant Microphones (MDM) in case of meeting recordings. The redundancy between different microphones contain information about speaker location. In this scenario, a beamforming algorithm [2] combines the different channels to a single better quality channel. This process selects a reference channel based on cross correlation with respect to other channels. Subsequently the Time Delay of Arrival (TDOA) of other channels with respect to this reference channel is computed for windowed segments of speech. The individual channels are then delay compensated and added together to obtain a single enhanced audio output from which 19 MFCC coefficients are extracted. The TDOA values contain location information about the speaker. Their combination with MFCC coefficients had achieved state of the art results in NIST diarization evaluations [3].

However, TDOA estimates are not accurate due to different variabilities such as reverberation, overlapped speech or background noise [4]. Furthermore, the global reference channel may not be the best quality channel with respect to different speakers with possibly changing locations. Hence, estimating the TDOA features with respect to only one reference channel can be suboptimal [4, 5]. There has been attempts in the past to employ time delay of arrival between all pairs of microphones to increase the robustness to such variations. However, the feature dimension of TDOA features depends on number of channels recorded and differs considerably across different meetings. The large dimension of features with all pairs of microphones has to be taken care separately. In [4] authors propose to select five best microphone pairs based on SNR and dynamic range. A histogram of TDOA features is built to perform a two stage clustering. The output of this stage is used as initialization to an MFCC based diarization. This method results in state of the art performance in NIST meeting diarization evaluations [6]. Alternatively, dimensionality reduction methods [5, 7] had been proposed for integrating large TDOA vectors as features for diarization. However, choosing the subspace dimension is not trivial. In [7] it was observed that the dimension that give minimum diariztaion error does not generalize well to the evaluation dataset.

This paper extends the authors previous work [8] where TDOA vectors from all pairs of features are used as features for diarization. In this work, it was noted that the performance of a HMM/GMM based diarization algorithm is sensitive to change of TDOA feature dimension. In spite of the global performance improvement, the performance deteriorates in case of several meetings in the evaluation data. In this work, the authors, investigate an unsupervised dimensionality reduction based on The Karhunen-Loève Transform (KLT) to circumvent the issue. The subspace dimension is selected by a scheme based on dimension of TDOA features in a distortionless recording. The resulting features are employed in combination with MFCC features for diarization.

The rest of the paper is organized as follows. Section 2 revises the TDOA feature estimation. The dimensionality reduction for the TDOA features is then presented in Section 3. Subsequently, the HMM/GMM system that is employed to perform speaker diarization is described in Section 4. Experiments and results on a combined dataset from various NIST evaluations is presented in Section 5. Finally the paper is concluded in Section 6.

2. TDOA Features

TDOA features are estimated across each possible pair of microphones using Generalized Cross Correlation with PHAse Transform (GCC-PHAT) [2]. The input audio recording is segmented into $500ms$ windows. Given signals corresponding to two channels, $x_i[n]$ and $x_j[n]$ the GCC-PHAT is defined by:

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

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

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

Table 1: Meeting number, identifier and associated number of microphones 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

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

Consider a meeting with M channels. When the delay features are estimated with respect to a reference channel, the resulting feature dimension would be $(M-1)$. We denote these features by *RTDOA*(Reference channel TDOA). On the other hand, when all the microphone pairs are considered for TDOA estimation the features would have a feature dimension of $M(M-1)/2$, which is much larger than the previous one. Henceforth these features will be referred as *ATDOA* (All pair TDOA).

3. Dimensionality Reduction

Dimensionality reduction aims to transform the high dimensional data into a low dimensional data with minimum loss of information. Linear dimensional reduction techniques such as KL transform [5], Unsupervised Discriminant Analysis (UDA) [7] had been employed in the context of TDOA features. In UDA, a set of local clusters are formed based on a nearness criterion and a subsequent discriminant analysis is performed. Since this require empirical thresholds about nearness of data, we consider a KLT for dimensionality reduction. The covariance matrix is estimated from the TDOA data samples $\{s_t^{tdoa}\}$ as:

$$C = \frac{1}{N}\sum_t (s_t^{tdoa} - m_t^{tdoa})(s_t^{tdoa} - m_t^{tdoa})^T \qquad (3)$$

where m_t^{tdoa} is the mean of the TDOA feature and N is the total number of frames. The set of K eigen vectors of C, that corresponds to K largest eigen values are selected, and projection to this subspace is used as features for diarization.

However, since the meetings are of variable dimension it is not apparent to choose K. It was observed that the optimum value of K changes considerably as the feature dimension changes. Eventhough different algorithms were suggested [9] (See chapter 5) to select the subspace dimension, none of them seems to significantly outperform a fixed dimensional subspace.

In this work we propose to choose the subspace dimension based on an idealistic situation. Consider the recording of a single sound source with no reverberation or environmental noise with a M microphone array. In this scenario, the signal recorded at each microphone $s_i[n]$ would be delayed and attenuated versions of a reference signal $s[n]$.

$$s_i[n] = A_i s[n - d_i] \qquad (4)$$

where A_i is the attenuation factor for channel i and d_i the amount of delay w.r.t reference signal $s[n]$. In this case the TDOA estimates would be given by $d_{PHAT}(i,j) = d_i - d_j$. Therefore there are only $(M-1)$ independent variables for this feature vector. If we perform KL transform on this vector we get zeros after $(M-1)$ dimensions. Motivated by this fact we assume that in a nearly idealistic condition of minimal environmental noise and overlap speech, most of the location information is contained in the first $M-1$ feature dimensions of the KLT. Thus we extract the first $(M-1)$ features corresponding to the largest eigen values of the TDOA covariance matrix as localization features. These features are referred as "STDOA-V"(Subspace TDOA with Variable dimension) henceforth. Note that this scheme does not involve selecting the feature dimension based on tuning on a development dataset.

4. Diarization System

The baseline diarization system is based on an ergodic HMM using a modified Bayesian Information Criterion [1, 10]. Each speaker is modeled with a HMM state with minimum duration. The emission probabilities are represented with GMM distributions. The emission probability distribution $b_{c_k}(s_t)$ of cluster c_k for input feature s_t is represented as $b_{c_k}(s_t) = \sum_r w_{c_k}^r \mathcal{N}(s_t, \mu_{c_k}^r, \Sigma_{c_k}^r))$, where $\mathcal{N}(.)$ denotes the Gaussian pdf and $(w_{c_k}^r, \mu_{c_k}^r, \Sigma_{c_k}^r)$ are the parameters of r^{th} mixture Gaussian.

The diarization system follows a bottom-up clustering. The system is initialized with a set of overestimated speaker clusters using uniform linear segmentation. Subsequently at each step a two clusters that are nearest according to a modified BIC criterion are merged together. Following each merge a viterbi realignment with the estimated speaker models(GMMs) is performed to refine speaker boundaries. The iterative merging stops when the BIC criterion across all pair of clusters are less than zero, thus determining the number of clusters.

4.1. Multiple Feature Input

When multiple features such as cepstral and TDOA features are available, the system builds individual GMMs for each feature stream. A linear combination of the individual log-likelihoods L_{c_k} is computed for each speaker cluster c_k

$$\log L_{c_k}(s_t) = w_{mfcc}\log[b_{c_k}(s_t^{mfcc})] + w_{tdoa}\log[b_{c_k}(s_t^{tdoa})] \qquad (5)$$

where w_i denotes the weight of feature stream i ($w_{mfcc} + w_{tdoa} = 1$). The combined likelihood $\log L_{c_k}(s_t)$ is used as the emission probability for the HMM system (replaces the term $\log b_{c_k}(s_t)$) in both clustering and realignment steps. Note that the log likelihood in Equation 5 depends on the TDOA feature dimension and thus the number of channels M. Therefore, the ATDOA features with feature dimension $(M-1)M/2$ is expected to have a large dynamic range of log likelihood as compared to STDOA-V features (dimension $M-1$).

Table 2: Dimension and Combination weights estimated for TDOA features. M denotes the number of microphones in the recording

Feature	dim	(w_{mfcc}, w_{tdoa})
ATDOA	$\frac{1}{2}M(M-1)$	$(0.999, 0.001)$
STDOA-V	$(M-1)$	$(0.9, 0.1)$
STDOA-F	7	$(0.9, 0.1)$

5. Experiments and Results

Experiments are performed on a dataset of 24 meetings collected across 6 meeting rooms . The set of meetings consists of all meetings from NIST RT'06/RT'07/RT'09 evaluations [11]. The set of meetings with number of channels is reported in Table 1. The multiple channels from the recording is beamformed using the *BeamformIt* [12] toolkit. 19 MFCC features are extracted from the beamformed audio, which are used in combination with the TDOA features.

The Time Delay of Arrival is estimated for each microphone pair. Computation of ATDOA features considers Only one array, in case of meetings recorded with multiple microphone arrays to avoid memory issues due to high feature dimension. Thus only the first 8 channels of EDI and IDI meetings are considered for ATDOA estimation. The performance is evaluated using Diarization error (DER) that is the sum of speech/non-speech detection and speaker error. Since the same speech non speech segmentation is applied across all experiments we report only the speaker error for comparison.

Dimensionality reduction is applied to the ATDOA features and the STDOA-V features are extracted as described in the previous section. For comparison purposes we also perform experiments to evaluate a fixed dimensionality reduction to a determined number of dimensions and the features are noted with "STDOA-F"(Subspace TDOA with Fixed dimensions). When this pre-determined value is higher than the TDOA dimension only the KLT is performed without any dimensionality reduction. The weights of feature combination in equation 5 as well as the dimension for STDOA-F are evaluated from a development dataset that consists of meetings from NIST RT05 evaluation. Note that in case of STDOA-F features the optimization involves searching a two dimensional grid of possible values for combination weights and dimension of the features.

Table 2 presents the estimated weights and estimated value for STDOA-F dimension. Whenever the the all pair TDOA features are used, the weights are quite different from the reference channel TDOA features as noted in the authors' previous work [8]. This phenomenon is related to the log likelihood combination (Equation 5). The ATDOA features have large dimension (eg: 28 for an 8 channel recording) and consequently, the TDOA log likelihood $\log[b_{c_k}(.)]$ increases. This leads to a quite different weight estimate for w_{tdoa} in case of ATDOA features. In case of STDOA features the weights are same as those estimated in terms of reference channel TDOAs [3]. Figure 1 shows the effect of tuning of weights for the two different subspace TDOA features. The dimensionality of both features is similar to reference TDOA features. In both cases the combination weights are same as in case of reference TDOA features.

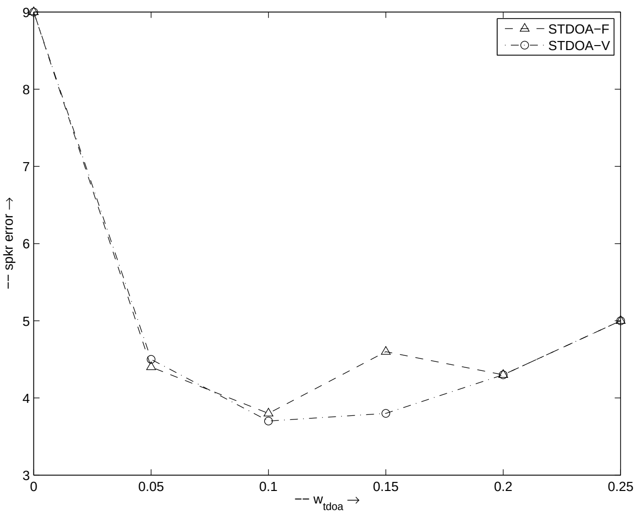

Figure 1: Speaker error value corresponding to different values of w_{tdoa}, $w_{mfcc} = 1 - w_{tdoa}$.The curve in case of STDOA-F represents the speaker error values with optimal feature dimension(7).

Table 3: Dimension and Combination weights estimated for TDOA features. M denotes the number of microphones in the recording

Feature	dev data	eval data
ATDOA	5.0	10.8
STDOA-F	3.8	7.5
STDOA-V	3.7	7.1

Results on development dataset (Table 3) indicate that the subspace TDOA features improve the speaker error by around 20% relative. Moreover, the STDOA-V features are able to attain similar performance as the STDOA-F features even though the latter is tuned for the best subspace dimension that minimizes the speaker error. Last column of Table 3 presents the results of the evaluation in test data. It can be seen that the dimensionality reduction helps to reduce the speaker error by around 3% absolute as compared to directly using ATDOA as features. This can be attributed to less variation in dimension of STDOA features that results in lower dynamic range for the likelihood values.

The two STDOA features give similar results for the dataset. Note that selected value of feature dimension for STDOA-F features is same as STDOA-V features for two of the NIST meetings as well as EDI and IDI meetings. Nevertheless, STDOA-V features result in a diarization error that is 0.4% (5%) relative than the fixed dimensionality reduction. Figure **??** illustrate the meeting-wise speaker error values for the ATDOA and STDOA-V features. Only in case of two meetings (meeting id 17,18), the performance drastically degrades. Further analysis show that the overlap speech is high in these meetings $(7 - 10\%)$ In most other meetings (17 out of 24s), dimensionality reduction of features show consistent improvement of performance. The STDOA-V features that do not require any tuning of subspace dimension outperforms the fixed dimension based STDOA-F features by around 5% relative.

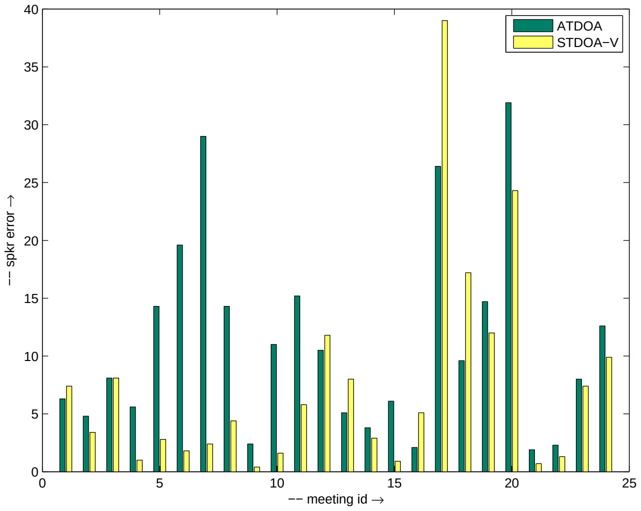

Figure 2: Meeting-wise speaker error values that correspond to ATDOA and STDOA-V features

6. Conclusions

Previous works in literature have investigated the use of TDOA features between all microphone pairs as features for diarization [5, 7, 8]. Issues related with increased dimensionality are addressed using different methods such as microphone pair selection or dimensionality reduction. In this work, we investigate a dimensionality reduction scheme to use all pairs of TDOA features into a diarization system. The number of features used selected after KL transform is made equal to one less than number of microphones. This is motivated from the fact that, in case of a microphone array recording under the conditions of minimal interference or environmental noise, the TDOA features would belong to a subspace of this dimension.

Experiments on a set of 24 meetings show that this scheme of choosing the subspace dimension reduce the dynamic range of likelihood across different meetings. The resulting features in combination with MFCC features improve the result by more than 3% (30% relative). This simple scheme has the advantage that it does not require any development data tuning to select features dimension. while retaining competitive performance.

It was observed that the performance of the method degrades when lot of overlap speech occurs in the recording (7-10%). Specialized algorithms to handle overlap needs to be investigated in this case (Example [13]). This would be addressed in the future research.

7. References

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

[2] 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.

[3] J.M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multi-microphone meetings: Mixing acoustic features and inter-channel time differences," in *International Conference on Speech and Language Processing*, 2006.

[4] 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.

[5] Scott Otterson, "Improved location features for meeting speaker diarization," in *Proceedings of Interspeech*, 2007.

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

[8] Deepu Vijayasenan and Fabio Valente, "Speaker diarization of meetings based on large tdoa feature vectors," in *Proceedings of ICASSP*, 2012.

[9] Scott Otterson, *Use of Speaker Location Features in Meeting Diarization*, Ph.D. thesis, University of Washington, 2008.

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

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

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

[13] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," 2008.