Selection of TDOA Parameters for MDM Speaker Diarization

Beatriz Martínez-González[1], José M. Pardo[1], Julián D. Echeverry-Correa[1], José A. Vallejo-Pinto[2], Roberto Barra-Chicote[1]

[1] Speech Technology Group, ETSI Telecomunicación, Universidad Politécnica de Madrid, Spain
[2] Department of Computer Science, University of Oviedo, Spain
{beatrizmartinez, pardo, jdec, barra}@die.upm.es, vallejo@uniovi.es

Abstract

Several methods to improve multiple distant microphone (MDM) speaker diarization based on Time Delay of Arrival (TDOA) features are evaluated in this paper. All of them avoid the use of a single reference channel to calculate the TDOA values and, based on different criteria, select among all possible pairs of microphones a set of pairs that will be used to estimate the TDOA's. The evaluated methods have been named the "Dynamic Margin" (DM), the "Extreme Regions" (ER), the "Most Common" (MC), the "Cross Correlation" (XCorr) and the "Principle Component Analysis" (PCA). It is shown that all methods improve the baseline results for the development set and four of them improve also the results for the evaluation set. Improvements of 3.49% and 10.77% DER relative are obtained for DM and ER respectively for the test set. The XCorr and PCA methods achieve an improvement of 36.72% and 30.82% DER relative for the test set. Moreover, the computational cost for the XCorr method is 20% less than the baseline.

Index Terms: Speaker diarization, speaker localization, speaker identification, speaker segmentation

1. Introduction

Speaker diarization is the task of identifying the number of participants in a meeting and creating a list of speech time intervals for each participant. Speaker diarization can be used as a first step in the speech transcription of meetings in which each sentence has to be associated with a specific speaker. The diarization task is carried out without any previous knowledge about the position, number or characteristics of the speakers, the position or quality of the microphones or the characteristics of the room where the recording has taken place. When the recording has been done with more than one distant microphone we speak of diarization with Multiple Distant Microphones (MDM).

Most MDM systems use acoustic features as Mel-Frequency Cepstral Coefficients (MFCC) and localization features as the Time Delay Of Arrival (TDOA) values [1]. Other features used in some systems are the normalized energy of the channels [2] or the prosodic parameters [3] [4].

The goal of this work is to improve the results of the diarization by improving or optimizing the TDOA values used in the segmentation and clustering. In [5] the baseline method to calculate TDOAs that is used in our system is described. It starts selecting one of the channels as the reference one (the channel with highest cross-correlations with the other channels) and estimating the TDOAs between this channel and the rest of them. The set of TDOAs from each microphone with the reference channel will form what we call the TDOA vector [tdoa] which, therefore, will have a dimension equal to the number of microphones minus one. This vector is used together with the MFCC vector in the subsequent segmenting and clustering procedure.

The aim of this work is to develop new methods to calculate the TDOA vector. In the current situation we are losing the possible information that the TDOA between any two microphones not selected could provide. The baseline system is using these two microphones to calculate the delay between each one of them with the reference microphone but not between themselves. We have tested five algorithms to select the microphones that we will use to calculate the TDOA vector. We have named these algorithms the "Dynamic Margin" (DM), the "Extreme Regions" (ER), the "Most Common" (MC), the Cross Correlation (XCorr) and the Principle Component Analysis (PCA). The explanation and the results of applying each method will be included in this work in the section with the same name.

Some works have already been carried out in this topic. In [6] all microphone pairs are used for computing correlation features. In [7] it has been presented a method to select the microphone pairs used to calculate TDOA's and the use of these TDOA's as the first stage of the segmentation and clustering module. Other alternative methods to select microphone pairs are presented in [8], [9], [10] and [11].

2. Database

In this work we have used a subset of 12 meetings extracted from NIST Rich Transcription 2002-2005 sets (named devel06 in [1]) and the RT06 and RT07 sets (from NIST Rich Transcription of years 2006 and 2007 respectively) to form our development set of 28 meetings that will be named ALL0607 from now on. The evaluation set will be the RT09 set, from the NIST Rich Transcription Evaluation of 2009.

The segments defined by NIST for the official evaluations have been used to measure the performance of the systems described in this work. In this paper we use the scored speaker time. These parts consist of 15,484.34 seconds (1,548,434 frames of 10 ms) evaluated for the ALL0607 set, and 5,932.88 seconds (593,288 frames of 10ms) for the RT09 set.

3. Baseline system

The input coming from several different microphones is first Wiener filtered in order to reduce the background noise. Then, in order to estimate the TDOA between two segments from two microphones, we use the Generalized Cross Correlation with Phase Transform" (GCC-PHAT). First, one of the channels is selected as the reference channel using the average cross-correlation between any pair of channels [12]. Then a TDOA value will be calculated every 250 ms for all the microphones with the reference one. For more detailed information see [5]. The set of TDOAs from each microphone to the reference microphone will form the TDOA vector [tdoa]. Once the [tdoa] vector is calculated, a weighted delay-

and-sum algorithm is applied in the acoustic fusion module, where the input signals are delayed and added together to generate a new composed signal. The composed signal is then processed by the MFCC estimation module, where MFCC vectors of 19 components [mfcc] are calculated every 10 ms with a window of 30ms. The composed signal is also processed by the Voice Activity Detector (VAD) module which is a hybrid energy-based detector and model-based decoder. The [tdoa] vector is also used in the subsequent module for clusters modeling but this time it is recalculated with a frame rate of 10ms in order to have the same number of data as the MFCC vector.

The algorithms presented in this paper modify this TDOA vector, thus, they affect the module for clusters modeling, but none of them are used to create the delayed and added signal necessary to extract the MFCC vector. That signal will be still calculated as in the baseline system and, consequently, MFCCs will remain unchanged in all the experiments.

The following module is the segmentation and agglomerative clustering process which consists of an initialization part and an iterative segmentation and merging process. The initialization process segments the speech into K blocks (equivalent to an initial hypothesis of K speakers or clusters) uniformly distributed. Every cluster is modeled using a gaussian mixture model (GMM) initially containing a number of components that has to be specified (we use 5 for [mfcc] and 1 for [tdoa] streams). After the initial segmentation a set of training and re-segmenting steps is carried out using Viterbi decoding. Then the merging step takes place.

When a merging takes place, the GMM for the new cluster is retrained with the data now assigned to it and the number of parameters (mixtures) of the merged model is the sum of the number of mixtures of the component models. The segmentation and clustering steps are repeated until a stopping criterion is reached. To decide which clusters to merge, and when to stop the merging, the BIC criterion has been used. When all possible merge pairs give a negative BIC, the merging is stopped. A frame purification algorithm is also applied before computing the BIC distance, see [12]. More information about the baseline system can be consulted in [1].

4. Methods for selecting delay features

4.1. Dynamic Margin

This method creates a histogram of delays for each possible pair of channels. These delays are calculated every 250ms along the whole recording. The histograms of TDOA values are generated ignoring the bins of the histogram with less than 25 samples. We use a bin width of 5ms. Then we select the subset of pairs with the highest dynamic range (highest difference between the maximum and the minimum delay). This method is very similar to the one presented in [7] although, in that paper, no details about the performance of the method were presented.

The optimum number of pairs has been chosen empirically after carrying out experiments from 1 to 10 pairs. The best performance was obtained for 3 pairs. Therefore, the method of selection will choose the 3 pairs with the highest dynamic range and then it will calculate the TDOA values for each frame of 10 ms, as it is done in the original method.

The DER obtained for the development set when using this method is shown in Figure 2 where values are given across different weights for the two streams of data: MFCC and TDOA. For most of the weights, the DER obtained with the DM method is better than the one with the baseline method. It is noticeable that one of the few points with worse performance is when only TDOA features are used (weight of MFCC stream equal to 0 in Figure 2). However, we do not intend to improve results in that point but in the area of lowest DER (around the baseline working point (MFCC weight=0.9)). The best result for DM method is obtained with the weight 0.85 for the MFCC stream and 0.15 for the TDOA stream. The baseline method obtains its best results with the weight 0.9 for the MFCC stream and 0.1 for the TDOA stream. Both values are shown in Table 1 where a slight but significant relative improvement can be seen for this method.

4.2. Extreme Regions

As it occurred with the DM method, the Extreme Regions method (ER) will begin calculating delays among all the different channels every 250ms. For each pair of microphones, all the bins with less than 25 appearances will be discarded. As in the previous method, we use a bin size of 5ms. Once this estimation has been carried out, the algorithm will make a histogram using all the values calculated from all the pairs of channels.

At this point some positions of the histogram are discarded. It has been decided to discard the most extreme positions that form 0.5% of the total number of delays calculated. This procedure aims to avoid some very high delays that are considered outliers. Straight afterwards two margins are set up, one in the positive part of the histogram and one in the negative part. These two margins define two regions which will contain 40% of the total number of delays calculated. In Figure 1 a hypothetical example of a histogram with the regions discarded (region A) and the regions with 40% of values (region B) is represented for better understanding.

Once defined the target region the system will check which of the pairs have more values in that area. As it happened in the dynamic range method, we chose only a subset of the total possible combinations. After carrying out experiments selecting a number of pairs from 1 to 10, the best performance has been obtained for only 2 combinations. The TDOA vector therefore, will have a dimension of 2. Once selected the 2 pairs the execution continues as usual, using these pairs to estimate the delay values for each frame of 10ms.

The DER obtained for the development set when using this method across different weights is shown in Figure 2. In this case the behavior of the ER method is worse than the baseline until the weight of MFCC streams turns higher than 0.4 when it starts to keep always below the baseline. As in the previous method, the best results have been obtained for a combination of weights of 0.85 for the MFCC stream and 0.15 for the TDOA stream. Results are shown in Table 1.

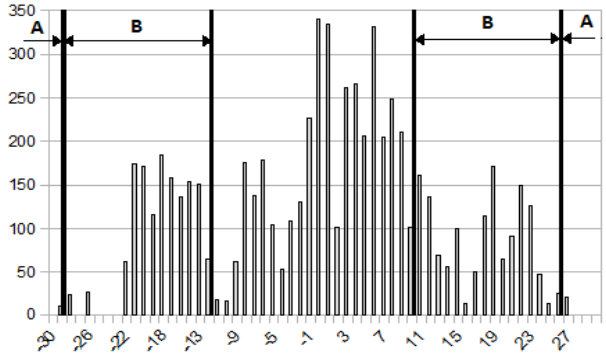

Figure 1: *Hypothetical example of the regions defined in the histogram of delays in the ER method. Region A is discarded (contains 0.5% of the total number of delays). Region B is target region (contains the 40% of the remaining number of delay values).*

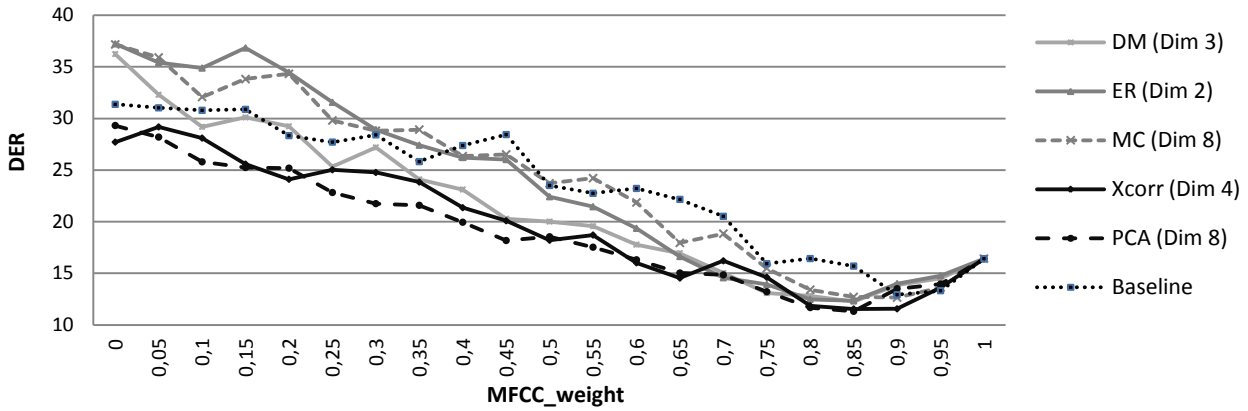

Figure 2: *DER of new methods for the development set (ALL0607), with their optimal dimension of the TDOA vector, using MFCC and TDOA features. Results are shown in compare to weight applied to the MFCC stream.*

4.3. Most Common

The third method presented in this paper will select first the most common TDOA values. Then, those pairs of microphones that have more TDOA values among those most common TDOA's, are the ones used in the final stage of clustering and segmentation.

This method starts, like the ER, calculating delays values for all the possible microphone pair combinations. All the information is joined in a histogram where each bin will correspond to a range of delay values. The bin width has been set to 5ms. The bins with highest values are selected as the most commons and the microphone pairs with the highest contribution to those bins are chosen.

To determine the best number of microphone pair combinations some experiments have been carried out. The results give the best performance for a number of combinations equal to 8. After the selection process the procedure is similar to those in previous sections.

The DER obtained for the development set when using this method is shown in Figure 2. The best numerical results obtained compared to the baseline are presented in Table 1. For this technique the best performance is obtained with the same weights than the best baseline system. This is MFCC weight equal to 0.9 and TDOA weight equal to 0.1.

4.4. Cross Correlation

As it was mentioned before, the baseline system selects one of the channels of the recording as the reference one. The system computes the average cross-correlation for all possible channel combinations for a block of duration 1s. This process is repeated for M=200 blocks linearly spaced along the recording. Then, it singles out the channel with the highest cross-correlation [12]. This channel has proved to be the most reliable channel to compute the delay with the other microphones. Using this kind of information we can select not one channel but several pairs of them. Using only pairs of channels with high cross correlation may avoid some bad estimation of delays and could lead to better performance of the system. For each pair of channels the cross-correlation is computed as:

$$xcorr_{i-j} = \frac{1}{M}\sum_{m=1,\mathrm{i}\neq\mathrm{j}}^{M} xcorr(i,j;m) \qquad (2)$$

Where M is the number of blocks used (M=200), and xcorr(i,j;m) indicates a standard cross-correlation measure between channels i and j for each block m.

The number of pairs finally selected, and therefore the dimension of the TDOA vector, has been set empirically to 4.

For this method and using 4 pairs to calculate TDOA values, we can see in Figure 2 that the DER keeps below the baseline for any MFCC weight. The lowest one, shown in Table 1, is obtained with a combination of weights for the MFCC and TDOA feature of 0.85 and 0.15 respectively.

4.5. Principle component analysis

With this method we try to take advantage of any information which can be held in any delay computed between any pair of channels. We intend to get as much information as possible of the delays. That would mean the computation of TDOA feature for every possible pair of channels. Then, we would perform a PCA to reduce dimensionality. However, some meetings have up to 24 microphones which means more than 250 combinations and a considerable increase in the computational time. To reduce this effect we have decided to use the cross correlation information we talked about in previous section. Now we are going to choose 50 pairs of channels using the XCorr method. Once the TDOAs between each pair have been found out, we will reduce dimensionality computing PCA.

Optimal final dimension of TDOA vector has been set empirically to 8. As it happened with the XCorr method the performance of the system is generally better than the baseline (see Figure 2). The best performance of this method is shown in Table 1. With this method, and weights for MFCC and TDOA features of 0.85 and 0.15 respectively, we obtained the lowest DER of all the methods implemented, although the difference between the Xcorr and PCA methods is small.

It is worth to mention that when using only TDOA features the results for the best methods, Xcorr and PCA, are better than the baseline (MFCC_weight equal to 0 in Figure 2), indicating that the TDOA information is more robust. This fact is also demonstrated at the optimum point in which the MFCC weight is 0.85 instead of 0.9 as in the baseline.

5. Evaluation

We mentioned in previous paragraphs optimization experiments to define the working point for each method. In Table 1 we included the best DER obtained for all the systems developed using the optimum parameters for the development set (ALL0607). The dimension of the TDOA vector calculated is included beside the name of each method. Weights for MFCC and TDOA features are 0.9 and 0.1 respectively for the baseline and the MC and 0.85 and 0.15 for the rest of them.

	DER for ALL0607	Improvement
Baseline	12.93 ± 0.05	
DM (Dim 3)	**12.24 ± 0.05**	**5.34%**
ER (Dim 2)	**12.36 ± 0.05**	**4.41%**
MC (Dim 8)	**12.65 ± 0.05**	**2.17%**
XCorr (Dim 4)	**11.55 ± 0.05**	**10.67%**
PCA (Dim 8)	**11.32 ± 0.05**	**12.45%**

Table 1: *DER for the set ALL0607 using MFCC and TDOA streams. Weight of MFCC stream for baseline and MC method is 0.9. Weight of MFCC stream for DM, ER, XCorr and PCA method is 0.85.TDOA stream weight is 1-MFCC weight.*

	DER for RT09	Improvement
Baseline	26.09 ± 0.11	
DM (Dim 3)	**25.18 ± 0.11**	**3.49%**
ER (Dim 2)	**23.28 ± 0.11**	**10.77%**
MC (Dim 8)	26.65 ± 0.11	-2.15%
XCorr (Dim 4)	**16.51 ± 0.09**	**36.72%**
PCA (Dim 8)	**18.05 ± 0.1**	**30.82%**

Table 2: *DER for the set RT09 using MFCC and TDOA streams. Weight of MFCC stream for baseline and MC method is 0.9. Weight of MFCC stream for DM, ER, XCorr and PCA method is 0.85.TDOA stream weight is 1-MFCC weight.*

In order to prove that the methods work properly for other sets of meetings we have evaluated the results with a new set of meetings: the RT09 set (7 meetings) using the parameters optimized for the development set. DER values for RT09 set are presented in Table 2.

While MC has not achieved any improvement for the test set, it can be seen that the results obtained for the evaluation set improve the baseline system for the DM, ER, XCorr and PCA methods. The XCorr and PCA methods clearly outperform the other two methods. In this case (with the test set) the improvement of the XCorr method is higher than the improvement obtained with the PCA method, and this time, the difference, is higher than what occurred with development set.

Finally we calculated the computational cost of all the systems developed. The results are shown in Table 3. Methods with the highest dimensions do not obtain any savings in time while those with lowest dimension reduce the computational time in about 20%. The PCA and XCorr methods are the best ones in DER for both the development and the test set. On average the XCorr has slightly better results than the PCA method but also, PCA method is much more costly due to the necessity of calculation of TDOAs for 50 pair of channels previous to the PCA dimensionality reduction. The XCorr method obtains both a great improvement in performance and a high reduction in computational time.

	Computational time of RT09
DM (Dim 3)	**0.77** *Baseline time
ER (Dim 2)	**0.82** *Baseline time
MC (Dim 8)	1.07 *Baseline time
XCorr (Dim 4)	**0.8** *Baseline time
PCA (Dim 8)	1.24 *Baseline time

Table 3: *Computational time for the set RT09 using MFCC and TDOA streams relative to computational time of baseline system.*

6. Conclusions

It has been shown that much better performance can be obtained using only a reduced set of values of TDOA when these have been selected properly. This paper has shown four ways of selection that achieve a relative improvement, over the test set, of 3.49% for the DM method, 10.77% for the ER, and 36.72% and 30.82% for XCorr and PCA methods respectively. The fifth method, MC method, worked well for the development set but not for the test set. Also, the computational cost is reduced around 20% with the three methods which use a TDOA vector of low dimension (DM, ER and XCorr). Computing TDOAs for a high number of channel combinations is computationally expensive, as it has been shown with PCA or MC method. Although, PCA has similar performance than XCorr method for the development set and the test set, the reduction in 20% in the execution time for the XCorr and the increase in 24% of this time for the PCA method, makes the XCorr the best option.

7. Acknowledgment

This work has been partially funded by projects TIMPANO (TIN2011-28169-C05-03) and INAPRA DPI2010-21247-C02-02 from the Ministry of Science and Innovation, and MA2VICMR, S2009/TIC-1542 by the Comunidad Autónoma de Madrid.

8. References

[1] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information,", *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1212–1224, Sept. 2007.

[2] R. Barra-Chicote, J. M. Pardo, J. Ferreiros, y J. M. Montero, "Speaker Diarization Based On Intensity Channel Contribution," *IEEE Transactions on Audio, Speech and Language* Vol 19, n 4, pp 754-761, May 2011.

[3] G. Friedland, O. Vinyals, Y. Huang, y C. Muller, "Prosodic and other Long-Term Features for Speaker Diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, nº. 5, pp. 985-993, 2009.

[4] J. M. Pardo, R. Barra-Chicote, R. San-Segundo, R. de Córdoba, B. Martínez-González, "Speaker Diarization Features: The UPM Contribution to the RT09 Evaluation*" IEEE Transactions on Audio, Speech and Language Processing*, Vol 20, No 2, pp 426-435, February 2012.

[5] X. Anguera, C. Wooters, y J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, nº. 7, pp 2011–2022, 2007.

[6] S. Otterson, "Improved location features for meeting speaker diarization," in *Proc. Interspeech*, págs. 1849–1852, 2007.

[7] H. Sun, T. L. Nwe, B. Ma, y H. Li, "Speaker diarization for meeting room audio," in *Proc. Interspeech*, pp. 900–903, 2009.

[8] T.L.Nwe, H.Sun, H.Li, S. Rahardja, "Speaker diarization in meeting audio", in *ICASSP 2009*, pp 4073-4076

[9] T.H.Nguyen, H. Sun et al, "The IIR-NTU Speaker Diarization Systems for RT 2009", *NIST RT09 Evaluation Workshop*, Florida, May 2009.

[10] N. Evans, C. Fredouille and J. -. Bonastre, "Speaker diarization using unsupervised discriminant analysis of inter-channel delay features," *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4061-4064, 2009

[11] D. Vijayasenan, F. Valente and H. Bourlard, "Mutual information based channel selection for speaker diarization of meetings data," *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4065-4068, 2009.

[12] X. Anguera. "Robust speaker diarization for meetings", *Ph D Thesis*, Universitat Politécnica de Catalunya, October 2006.