AUTOMATIC WEIGHTING FOR THE COMBINATION OF TDOA AND ACOUSTIC FEATURES IN SPEAKER DIARIZATION FOR MEETINGS

Xavier Anguera[1,2], *Chuck Wooters*[1], *Jose M. Pardo*[1,3] *and Javier Hernando*[2]

[1] International Computer Science Institute, Berkeley, CA 94704, U.S.A.
[2] Technical University of Catalonia (UPC), 08034 Barcelona, Spain
[3] Universidad Politecnica de Madrid, 28040 Madrid, Spain
{xanguera,wooters}@icsi.berkeley.edu

ABSTRACT

In the task of speaker diarization for meetings it has been shown in previous work that it is useful to use the Time Delay of Arrival (TDOA) between the different audio channels in the meeting room as an extra source of information in addition to the acoustic features. When combining feature streams, we use a weight to control the relative contributions of the streams. In the past, this weight was determined using development data and the same weight value was applied to all meetings. In this paper we present a method for automatically determining the weight. A metric derived from the Bayesian Information Criterion (BIC) computed for each feature stream estimates the weight for each meeting on the initial clustering iteration and adapts its value throughout the diarization process. By using this technique we achieve a more robust system and up to 18.2% relative improvement over the method of tuning the weight on development data.

Index Terms— Speaker diarization, segmentation, clustering, BIC, features fusion, multi-stream

1. INTRODUCTION

The task of speaker diarization involves the automatic segmentation and clustering of acoustic data into speaker homogeneous regions, attempting to answer the question "who spoke when?". Speaker diarization is usually performed without any prior information regarding the number of speakers or their identities. The most common technique used for this task is "bottom-up" agglomerative clustering, which first splits the acoustic data into a large number of clusters and then iteratively merges pairs of clusters until a stopping criterion indicates that the merging should stop.

Xavier Anguera and Jose M. Pardo were visiting ICSI within the Spanish visitors program at the time of this work. Xavier Anguera has been partially supported by project TIN-2005-08852.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

In the meeting domain, we usually have access to multiple microphones that synchronously capture the audio in the meeting. The use of the Time Delay of Arrival (TDOA) between the microphones for speaker diarization has been used in the past, either independently [1], [2] or in combination with acoustics [3], [4]. Independently of the method used for the combination of these two feature streams, a weighting between them needs to be applied to take clustering and segmentation decisions.

In [4], [5] we proposed a system that obtains the TDOA values by applying an acoustic beamforming to all available channels and then combines it with the acoustic features by a weighted sum at the log likelihood level. The weights needed to be tuned by hand using development data and was fixed for all meetings. This imposes a restriction on the number of different features to use, as the search space grows exponentially with the number of streams used. It is also does not adapt to each meeting type, which can alter the optimum weight to be used.

In this paper we describe a way to automatically determine the weight between acoustic and TDOA features on a per-meeting basis, which can be extended to as many feature streams as desired. Previous work in weight selection for feature fusion has to be looked for in areas other than speaker diarization, such as in speaker verification and biometrics [6], [7] and in speech recognition [8], [9]. A common technique for automatic weighting of different feature streams is based on the inverse entropy of the classes predicted by the feature vectors. This is not useful when combining TDOA values and acoustic features because of the particular characteristics of the TDOA pdf function, which often considers non-existent speaker locations with a high probability.

The proposed technique uses a metric derived from the Bayesian Information Criterion (BIC), used as a cluster comparison metric to compare how well each feature stream differentiates between clusters. It is first computed after the initial ΔBIC metric between each cluster pair and it is adapted after each consecutive iteration obtaining weights that converge over time.

2. AGGLOMERATIVE SPEAKER DIARIZATION SYSTEM

The agglomerative speaker diarization system used in this paper is shown in figure 1. It is based on the system used in the Rich Transcription evaluations on meetings (RT06s) as described in [5]. The signals from the multiple available microphones are first analyzed by a filter&sum beamforming [10] in order to obtain a single enhanced channel. Independent feature streams are then created from the acoustic data (19 MFCC features computed every 10 msec) and from the TDOA values.

The input acoustic signal is then processed by a speech/non-speech detector to eliminate the non-speech regions from the clustering process. Such detector uses a hybrid energy/model-based approach in a semi-supervised manner. Then models are trained from the initial set of clusters, one for the acoustic stream and one for the TDOA values for each cluster. In the current implementation these models contain 5 mixtures for the acoustics and a single Gaussian for the TDOA values. The combination of both streams is done at both the segmentation and clustering stages. At the segmentation stage, the joint log-likelihood for any given frame is computed as

$$\mathcal{L}(x_{aco}[n], x_{del}[n]|\Theta_{aco}, \Theta_{del}) = \mathcal{W}_1 \cdot \mathcal{L}(x_{aco}[n]|\Theta_{aco}) + \mathcal{W}_2 \cdot \mathcal{L}(x_{del}[n]|\Theta_{del}) \quad (1)$$

where Θ_{aco}, $x_{aco}[n]$ is the acoustic model and acoustic data, Θ_{del}, $x_{del}[n]$ is the TDOA model and TDOA data, and $\mathcal{W}_1 + \mathcal{W}_2 = 1$ weight the effect of each model in the system. It is the estimation of $\mathcal{W}_i$ that is the focus of this paper. In this formulation we consider the streams to be statistically independent from each other.

At the clustering stage, a modified version of the Bayesian Information Criterion (BIC) is used (see [11] as a cluster-pair distance metric and as a stopping criterion. The combination of both feature streams in the clustering stage is done with

$$\Delta BIC(A, B) = \mathcal{W}_1 \Delta BIC_{acous}(A, B) + \mathcal{W}_2 \Delta BIC_{del}(A, B) \quad (2)$$

where A, B are the two clusters we are comparing, and $\mathcal{W}_i$ is the same weight as in eq.1.

The system iteratively resegments the data using eq. 1 and computes the closest cluster pair using eq. 2 while $\Delta BIC(A, B) > 0$ for any pair A, B.

3. STREAM WEIGHT SELECTION ALGORITHM

As seen in equations 2 and 1, in order to combine the acoustic and TDOA features one needs to determine the value of $\mathcal{W}_i$, which specifies how much relevance is given to each stream. Setting the values of $\mathcal{W}_i$ by hand can lead to a robustness problem due to differences between development and test sets. Furthermore, when setting the value experimentally, we typically use a single value for all meetings and therefore it can not account for peculiarities of the individual meeting rooms (noisier rooms) or of the nature of the meetings (type of attendees or wether they move from their seats). Finally, manual tuning becomes unfeasible if the number of feature streams is big (where all streams are combined using a weighted sum as in equation 1).

There are many techniques for performing acoustic fusion of multiple feature streams. A common technique is based on entropy. Initial tests were performed using the inverse entropy as relative weight to see how discriminant each feature stream was. This was done by obtaining the weights in a frame-basis via the inverse entropy of the posterior probabilities of the cluster models given the data. For MFCC, PLP and other acoustic features these entropies were comparable to each other and could therefore determine a correct relative weight between features, as shown in [8]. When using it with TDOA values their GMM models are such that low entropy values are obtained for almost every frame, regardless of how accurate the TDOA values can represent a real speaker position.

The proposed technique in this paper uses the Bayesian Information Criterion (BIC) to compare how well each feature stream differentiates between clusters in order to determine an appropriate stream weighting. The ΔBIC values are independent of the complexity and topology of the models being used and are a good indication of how close two clusters are. In order to allow for different feature streams to contribute equally in the merging decision it is needed to transform both ΔBIC value sets to have the same scale using the $\mathcal{W}_i$ weight. This way the TDOA values with overall high ΔBIC are penalized versus the acoustic values in order to be comparable to each other. For a general case of M feature streams, the weight $\mathcal{W}_i$ assigned to each stream i is defined as

$$\mathcal{W}_i = \frac{\frac{1}{\sqrt{P_i}}}{\sum_{j=1}^{M} \frac{1}{\sqrt{P_j}}} \quad (3)$$

where P_i is computed from the N ΔBIC values computed for all cluster pairs x_j, x_k from each feature stream as

$$P_i = \frac{1}{N} \sum_{j=1}^{j=N-1} \sum_{k=j+1}^{k=N} \Delta BIC_i^2(x_j, x_k) \quad (4)$$

This process is equivalent to a variance normalization of single Gaussians modeling each feature stream with zero mean. Setting the mean to zero avoids moving the decision threshold in the ΔBIC comparison, as defined by the BIC theory.

The automatic computation of the $\mathcal{W}_i$ weight is performed at the first clustering step, when the ΔBIC values are computed. At the initial segmentation step, no weight has been automatically defined and therefore some initial weight still

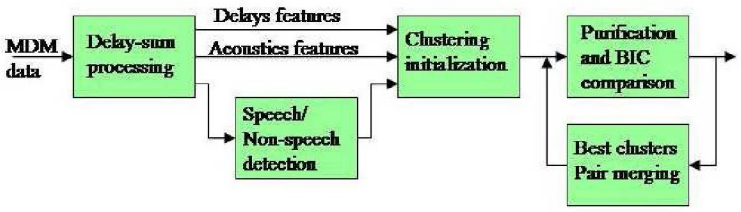

Fig. 1. *ICSI speaker diarization system for multichannel recordings*

needs to be determined by hand, or it can be set to an uninformative $\mathcal{W}_1 = \mathcal{W}_2 = 0.5$. On subsequent clustering iterations the models usually represent the clusters better and obtain ΔBIC values which are more accurate. In order to allow the system to refine the weight as the merging iterations progress, the ΔBIC values are kept for all cluster pairs that disappeared during previous iterations and existing pairs are recomputed. Then a new weight is computed taking into account both old and updated values in order to allow for a weight adaptation, containing enough samples for a robust computation.

4. EXPERIMENTS

In order to test the effectiveness of the proposed automatic weighting scheme, we use the current ICSI speaker diarization system as described in section 2 without the use of any purification [5] and using linear clusters initialization.

The development data is composed of the data sets prepared for the NIST Rich Transcription (RT) evaluations ([12]) used for RT02, RT04s and RT05s for conference data, excluding those meetings with only one channel (where no TDOA values can be computed). This forms a set with 22 meeting excerpts with durations of 10-12 minutes each and containing different numbers of speakers and meeting rooms. As test data we use the set used for the RT06s evaluation, consisting of 8 meeting excerpts with an average duration of 15 minutes each. The metric used in all cases is the Diarization Error Rate, defined by NIST [12] as the percentage of misassigned time.

The algorithm performance is compared to the same system using only the acoustic features (equivalent to assigning $\mathcal{W}_1 = 1$), and to the multistream version where the stream weight was determined based the development data and set to $\mathcal{W}_1 = 0.9$ for all meetings.

One possible parameter in the new algorithm is the number of iterations in which the weights are to be recomputed. To illustrate the effect of the weight adaptation as the system iterates, figure 2 shows the DER of the development set and the weight evolution of the show CMU_20050912-0900 (chosen randomly from the test set) from 1 to 10 iterations of automatic weights computation. The DER decreases as the number of iterations increase, with the exception of iteration 3, stabilizing around iteration 9. This indicates that the system tends to obtain better values for the weight as it progresses, and therefore there is no need to tune the number of

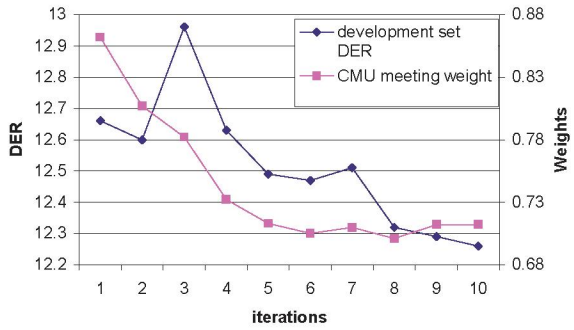

Fig. 2. *Weights and DER evolution with the weight computation iterations*

iterations. Instead, we allow it compute a new weight as long as the stopping criterion does not stop the system.

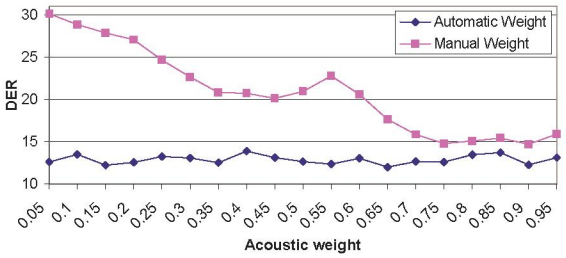

Fig. 3. *Comparison of automatic versus manual weights setting*

In order to run the initial segmentation a first value for the $\mathcal{W}_i$ weights must be chosen. Figure 3 shows the DER on the development data set comparing the effect of this initial selection on the proposed system versus setting the same weight in the standard algorithm (without automatic selection). Although a slight ripple is seen in the automatic algorithm performance depending on this initial weight, it is small compared to the effect seen in the manual weight curve. The proposed system performs correctly for any initial guess in the streams weight. In a real application a rough initial weight is enough to initialize the system

In table 1 we compare the DER of several implementations. The mono-stream system uses only acoustic features, the other systems use both acoustics and TDOA values, differing in the way that the weights are found. The system

"inv-entropy" performs a frame-wise inverse entropy weight estimation as described in [8]. The "manual weights" system finds the optimum weights using a development set and is set to $\mathcal{W}_1 = 0.9$. The other two lines show results using the automatic weighting with different initial weights, $\mathcal{W}_1 = 0.9$, optimum in the development set for manual case and $\mathcal{W}_1 = 0.65$ optimum in the development set for the automatic weight setting.

System	weight	DER Devel	DER Test
mono-stream	n/a	17.65%	26.50%
inv-entropy	auto	24.94%	28.57%
manual weights	optimum(0.9)	14.7%	**18.65%**
auto-weights	0.9 + auto	12.28%	20.07%
auto-weights	opt(0.65) + auto	**12.01%**	20.87%

Table 1. DER results for for different weight setting algorithms

At a first glance we see that using inverse entropy does not achieve good results. In average the entropy method assigns higher weight to the TDOA values while all optimum performance points do otherwise. Also, observe that all the multistream methods (except inv-entropy) greatly outperform the monostream system.

Automatic weighting obtains, in its optimum point, a relative 18.2% improvement versus manual weighting in the development set. Manually setting the weight achieves the best performance in the test set, although this is misleading as values around that weight obtain much higher errors (DER = 22.85 for $\mathcal{W}_1 = 0.85$ and DER = 22.29 for $\mathcal{W}_1 = 0.95$) which makes us doubt of its performance in other data sets. In addition, the values for the automatic weighting algorithm in the test set remain stable (DER = 20.5% in average) for most observed weights.

Even though performance is slightly worse than the baseline in the test set, we believe on the capability of the algorithm to automatically find the feature stream weights without the need of tuning any parameter. This becomes a key issue if more than two features are used in the system, in which manual tuning becomes more difficult to perform.

5. CONCLUSIONS

In this paper a novel technique is presented for automatic weight estimation for multistream speaker diarization for meetings. When multiple microphones are available for processing it has been shown that using the TDOA information helps immensely the speaker diarization via a weighted sum with the acoustics at the likelihood level. Standard techniques as inverse entropy are found not successful in automatically define the weights between features. The method proposed is based on the equalization of a metric obtained from the ΔBIC values computed for cluster pairs and it is shown as a feasible alternative to manually setting the weights. Improvements of up to 18.2% relative are shown on the development set.

6. REFERENCES

[1] D.P.W Ellis and Jerry C. Liu, "Speaker turn detection based on between-channels differences," in *Proc. ICASSP*, 2004.

[2] G. Lathoud, I.A. McCowan, and J.M. Odobez, "Unsupervised location-based segmentation of multi-party speech," in *ICASSP-NIST Meeting Recognition Workshop*, May 2004.

[3] Jitendra Ajmera, Guillaume Lathoud, and Iain McCowan, "Clustering and segmenting speakers and their locations in meetings," in *Proc. ICASSP*, 2004, vol. 1, pp. 605–608.

[4] Jose M. Pardo, Xavier Anguera, and Chuck Wooters, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences," in *Proc. ICSLP*, September 2006.

[5] Xavier Anguera, Chuck Wooters, and Jose M. Pardo, "Robust speaker diarization for meetings: ICSI RT06s evaluation system," in *Proc. ICSLP*, Pittsburgh, USA, September 2006.

[6] J. Fiérrez-Aguilar, J. Ortega-García, and J. González-Rodríguez, "Fusion strategies in multimodal biometric verification," in *IEEE International Conference on Multimedia and Expo*, 2003.

[7] A. Ross, A. K. Jain, and J. Z. Qian, "Information fusion in biometrics," in *3rd International Conference on Audio and Video-Based Person Authentication*, 2001.

[8] Hemant Misra, Herve Bourlard, , and Vivek Tyagi, "New entropy based combination rules in hmm/ann multi-stream asr," in *Proc. ICASSP*, Hong Kong, 2003.

[9] S. Ikbal, H. Misra, S. Sivadas, H. Hermansky, , and H. Bourlard, "Entropy based combination of tandem representations for noise robust asr," in *Proc. ICSLP*, South Korea, 2004.

[10] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *Proc. ASRU*, Puerto Rico, USA, November 2005.

[11] Jitendra Ajmera and Chuck Wooters, "A robust speaker clustering algorithm," in *Proc. ASRU*, US Virgin Islands, USA, Dec. 2003.

[12] "NIST rich transcription evaluations, website: http://www.nist.gov/speech/tests/rt," 2006.