Friends and Enemies: A Novel Initialization for Speaker Diarization

Xavier Anguera[1,2], *Chuck Wooters*[1], *Javier Hernando*[2]

[1] International Computer Science Institute
1947 Center St., Suite 600
Berkeley, CA 94704, U.S.A.
[2] Technical University of Catalonia (UPC)
Jordi Girona 1-3, building D5
08034 Barcelona, Spain
{xanguera,wooters}@icsi.berkeley.edu

Abstract

The task of speaker diarization consists of answering the question "Who spoke when?". The most commonly used approach to speaker diarization is agglomerative clustering of multiple initial clusters. Even though the initial clustering is greatly modified by iterative cluster merging and possibly multiple resegmentations of the data, the initialization algorithm is a key module for system performance and robustness. In this paper we present a novel approach that obtains a desired initial number of clusters in three steps. It first computes possible speaker change points via a standard technique based on the Bayesian information criterion (BIC). It then classifies the resulting segments into "friend" and "enemy" groups to finally creates an initial set of clusters for the system. We test this algorithm with the dataset used in the RT05s evaluation, where we show a 13% Diarization error rate relative improvement and a 2.5% absolute cluster purity improvement with respect to our previous algorithm.

Index Terms: Speaker diarization, speaker segmentation and clustering, clusters initialization, meetings indexing.

1. Introduction

The goal of speaker diarization is to segment an audio recording into speaker-homogeneous regions [1] answering the question "Who spoke when?". Typically, this segmentation must be performed with little knowledge of the characteristics of the audio or of the participants in the recording. We can normally know the type and source of the recording (wether it is a meeting or broadcast news, and when it was recorded). We cannot use any information on the number of speakers present, their identities, noise, commercials or other events.

The most commonly used technique in speaker diarization is based on agglomerative clustering. An initial set of clusters is iteratively reduced by merging the closest pair according to a similarity metric until a stopping point is reached. A cluster is defined to be a set of segments, not necessarily contiguous, that share some acoustic similarity. A segment is defined to be a contiguous set of acoustic frames. In the system presented here we constrain the segments to have a minimum duration. It is also common practice to use BIC [2] as a similarity metric between clusters and as a stopping criterion.

In order to define the initial clustering we need to establish a tradeoff between simplicity and accuracy. It can be thought that such initialization is of small importance given the multiple iterations of resegmentation and merging of the data that are performed in the clustering process. A simple linear initialization of the data into the desired number of equal-sized clusters has been used with relative success in our speaker diarization system. Acoustic models are trained from such data and a resegmentation-retraining process is used to redistribute the data into homogeneous clusters. This method is very quick and brings relatively good results.

By using linear initialization, no constraint is applied to the data that is initially categorized into each cluster, leaving it to the resegmentation and retraining process to reassign the acoustic data into clusters. In some cases, acoustic segments from more than one speaker remain in their originally-assigned cluster throughout the clustering process. In other cases a minority speaker remains within a cluster containing data from another speaker and ends up merging with this speaker. In both situations we end up with an increase in the diarization error rate (DER) which is difficult to reduce during the clustering process. The use of a standard k-means algorithm (see [6])to find the cluster initialization obtains even worse results than with the linear initialization algorithm.

In this paper we present a novel initialization algorithm that aims at creating an initial clustering with a predefined number of clusters with emphasis on cluster purity. We use the definition of purity introduced in [3], which accounts for the percentage of frames in any given cluster that come from the most represented speaker in that cluster. It differs from the DER in that we don't try to find the optimum number of clusters (we would obtain perfect purity if a different cluster was created for each frame of data). To do so, we first find the most probable speaker change points in the recording using the BIC metric. Then we make groups of friends using these segments until reaching the desired number of initial clusters. Finally we reassign all frames to the newly all created clusters.

In section 2 we review the speaker diarization system used in this paper. In section 3 we present the proposed algorithm and in section 4 we show experiments comparing this algorithm to the previously used one. Finally we draw some conclusions.

2. Agglomerative Speaker Diarization System

As explained in [4], [5] and [6], the speaker clustering system is based on an agglomerative clustering technique. It initially splits the data into K clusters (where K must be greater than the num-

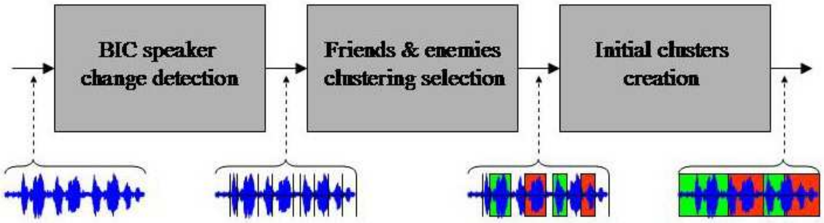

Figure 1: *Clusters initialization blocks diagram*

ber of speakers, and is chosen using the algorithm presented in [7]), and then iteratively merges the clusters (according to a metric based on ΔBIC) until a stopping criterion is met. Our clustering algorithm models the acoustic data using an ergodic hidden Markov model (HMM), where the initial number of states is equal to the initial number of clusters (K). Upon completion of the algorithm's execution, each remaining state is taken to represent a different speaker. Each state in the HMM model contains a set of MD sub-states, imposing a minimum duration on the model (we use $MD = 3$ seconds). Within the state, each one of the sub-states shares a probability density function (PDF) modelled via a Gaussian mixture model (GMM).

The system works as follows:

1. If more than one recorded channels is available for a given meeting, we combine them into a single "enhanced" channel using a delay-and-sum algorithm [8].
2. Run speech/non-speech detection on the enhanced channel using the speech/non-speech algorithm presented in [9].
3. Extract acoustic features from the data and remove non-speech frames.
4. Estimate the number of initial clusters K using the algorithm presented in [7].
5. Create models for the K initial clusters using either linear initialization or the new proposed initialization algorithm.
6. Perform iterative merging using the following steps:
   (a) Run a Viterbi decode to resegment the data.
   (b) Retrain the models using Expectation-Maximization (EM) and the segmentation from step (a). Repeat steps (a) and (b) several times to stabilize the segmentation.
   (c) Select the cluster pair with the largest merge score (based on ΔBIC) that is > 0.0.
   (d) If no such pair of clusters is found, stop and output the current clustering.
   (e) Merge the pair of clusters found in step (c). The models for the individual clusters in the pair are replaced by a single, combined model.
   (f) Go to step (a).

For the merging distance measure and clustering stopping criteria, we use a variation of the commonly used Bayesian Information Criterion (BIC) [10]. The ΔBIC compares two possible models: two clusters belong to the same speaker or they belong to different speakers. The variation we use was introduced by Ajmera et al. [6], [11], and consists of the elimination of the tunable parameter λ by ensuring that, for any given ΔBIC comparison, the difference between the number of free parameters in both models is zero.

The cluster initialization block is often been considered to be of less importance, as many segmentations and model retraining iterations take place later in the process that would allow a sub-optimal initialization to perform as well as any other. In this respect it has been considered that the best initialization is that which doesn't introduce any computational burden to the overall system. With a marked reduction of the error in the current system, we have seen that the linear initialization *does* cause a problem on the final score, since some initial clustering errors are propagated all the way to the end of the agglomerative clustering and show up in the final result. It has also been seen that a linear initialization without any acoustic constraints on the created clusters introduces a random effect in the system which could be one of the sources of per-show "flakiness", as presented in [12].

When designing an initialization algorithm for speaker diarization there is an additional problem beyond the standard problem of acoustic clustering. It is important to constrain the classification of the acoustic information according to its acoustic context, as it will be afterwards classified within the rest of the system, which uses a minimum duration for a speaker turn to avoid instabilities and very short segments. For this reason it is important to split into separate initial clusters acoustic data from a speaker in different background situations (for example in a solo presentation, in overlap or between many non-speech events).

In the next section we present the proposed cluster initialization algorithm that addresses these problems, while not imposing a significant burden on the system's speed, and then we compare it to the standard linear initialization in section 4.

3. Friends Versus Enemies Initialization Algorithm

The proposed initialization algorithm is designed to split the acoustic data into N clusters, where N is determined beforehand by some other algorithm or set by the user. In the agglomerative clustering scheme presented here, N corresponds to the initial number of clusters used to start the agglomerative process. Each of the resulting initial clusters has a duration which is not restricted to be equal to any other cluster.

The complete initialization is composed of three distinct blocks, as shown in Figure 1. The first block performs a speaker-change detection on the acoustic data to identify segments with a high probability of containing only one acoustic event. Such acoustic events can be silence, various noises, an individual speaker or various speakers overlapping each other. We perform this first step using the modified Bayesian Information Criterion (BIC) metric (introduced by [6]) computed between two models

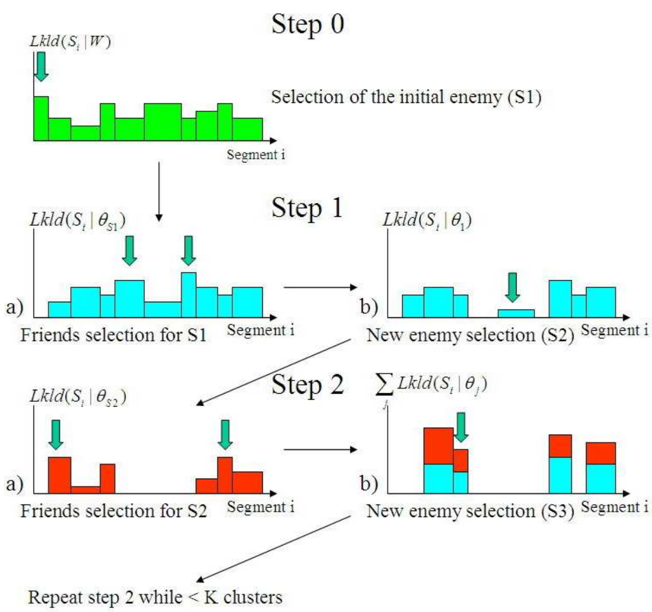

Figure 2: *friends vs. enemies clusters initialization process*

created from the data in two adjacent windows of size W, connected at the considered change point. The modified BIC metric is computed over all the acoustic data every S frames. A possible change point is selected if $BIC < 0$ and it corresponds to a local minimum of the BIC values around it.

The second block creates clusters by identifying the segments defined in the first part as friends or enemies of each other. We consider that two acoustic segments are friends if they contain acoustically homogeneous data; only the best friends are brought together to form a cluster. In the same way, we consider two segments to be enemies if they contain very dissimilar acoustic data. Our aim is to obtain N final enemy groups (the desired final number of clusters) consisting of F segments each, which are friends of each other.

We can see in figure 2 how the algorithm works. Given all the acoustic data to be processed, we build a general model W with 16 gaussian mixtures. The top left graph shows the cross-likelihood of each segment S_i given the world model W normalized by the number of frames in each segment. The segment with the lowest normalized cross-likelihood ($\overline{xlkld}$), with the world model, is taken as the initial cluster/enemy S_1. The expression used for the $\overline{xlkld}$ is

$$\overline{xlkld}(S_1, S_2) = \frac{lkld(S_1|\Theta_{S2}) + lkld(S_2|\Theta_{S1})}{(L_{S1} + L_{S2})} \tag{1}$$

where L_{S1} and L_{S2} are the length of segments S_1 and S_2 respectively.

In step 1a in figure 2 we use the data in S_1 to train a model with 5 gaussian mixtures (Θ_{S1}) and compute the $\overline{xlkld}$ with all other segments. The $F-1$ segments with bigger $\overline{xlkld}$ are its friends. In this example, $F = 3$. In step 1b, a new model is trained from all data in this group (Θ_1) and the $\overline{xlkld}$ with all remaining segments is computed. A new enemy S_2 is also selected as the segment with smaller $\overline{xlkld}$. Also in the same way, in step 2a we select $F-1$ friends for S_2 and in 2b we select a new enemy for both previously established clusters. This is done by computing the sum of the $\overline{xlkld}$ for each segment given all predefined groups. The processing continues until the desired number of initial clustering N is reached or we run out of free segments.

At that point in the third block we use all created models to reassign the acoustic data into the N classes using Viterbi. The resulting clustering is not constrained to the predefined speaker changes, therefore any speaker change detection errors can be corrected. All data gets assigned to its closest cluster, classifying any acoustic frames not assigned in the previous block. Finally, one cluster model is trained from each of the resulting clusters.

4. Experiments

In order to test the proposed initialization algorithm, we compare its performance to the linear initialization used in our speaker diarization system to date. Such initialization defines N initial clusters by splitting the input signal into even parts and then iterates over model training and segmentation on the data in order to obtain initial clusters with acoustically homogeneous data.

Both initialization techniques were compared using the data distributed for the NIST Rich Transcription 2005 Spring Meeting Recognition Evaluation, RT05s ([13]). This consists of excerpts from multi-party meetings in English collected at six different sites at various time periods. From each meeting only an excerpt of 10 to 12 minutes is evaluated. Varying numbers of microphones are available for each recording ranging from 3 to 16. We processed all available microphones using an implementation of the delay-and-sum algorithm (see [8]) to obtain a single enhanced signal, on which we apply the diarization algorithms.

In order to compare the two techniques, we measure their performance at two different stages of the speaker diarization system: cluster purity and diarization error rate (DER).

We compute a cluster's purification right after the initialization algorithm. We use the concept of purity as introduced by [3], where for each initial cluster we compute the percentage of the total cluster time used by the main speaker present in that cluster according to the reference clustering. The total cluster purity for a particular recording is the time-weighted sum of all individual cluster purities. In the same way, the overall cluster purity is the time-averaged sum of all individual recording purities. A cluster purity of 100% indicates that all clusters contain only one speaker.

In addition, we use the diarization error rate (DER) as used in the NIST Rich Transcription Evaluations, to measure the overall diarization score. DER is computed by first finding an optimal one-to-one mapping of reference speaker ID to system output ID and then obtaining the error as the percentage of time that the system assigns the wrong speaker label. It differs from the cluster purity in that it looks at the overall accuracy. As with the purity metric, the time-weighted DER score is reported for the group of meetings in each evaluated set.

The results for the proposed algorithm were obtained using the following parameters: for the speaker change detection step, individual windows of two seconds were used, with the BIC metric computed every half a second. Change points are allowed only when the distance between any two change-points is greater than three seconds. For the friends and enemies block, we used five gaussian mixtures per cluster. Figure 3 shows the cluster purity and DER for the RT05s set using different values for F. At $F = 3$ both cluster purity and the DER have their optimum values, although it is not clear whether a better cluster purity always correlates with a lower final DER. There are other points in the clustering pro-

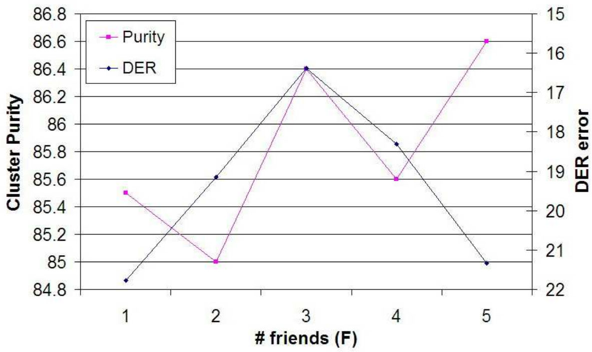

Figure 3: *Cluster purity and DER for different values of F*

cess (cluster comparison, stopping criterion, etc) that can impact negatively on the final DER.

The results on cluster purity and DER for both compared systems are shown in table 1.

Initialization system	Cluster purity	DER
Linear init.	83.9%	18.82%
Friends/enemies init.	**86.4%**	**16.38%**

Table 1: Cluster purity and DER for the alternative initializations

We obtain an improvement of about 13% relative in the DER by using this new initialization, with an improvement of 2.5% in the cluster purity right after the initialization algorithm. The large differences obtained in the final DER by using different initialization techniques indicate how important is it to obtain an accurate representation of the speakers in order to have accurate speaker diarization using the agglomerative clustering approach. By these results, comparing both DER and cluster purity, we see that it is apparently important to design algorithms that have a high purity at early stages, but it is not clear if this is the only requirement we should impose on the initial clusters in order to obtain a better DER. In the meetings environment there is a significant amount of overlap speech which should be taken into account when creating the initial clusters, as it is very likely that such overlap segments will affect the clustering decisions and therefore the final result. While using the algorithm presented here, it is likely that overlap speech will be assigned its own cluster, so cluster purity alone might not be suitable for measuring how well it performs.

5. Conclusions

In this paper we presented a novel algorithm for cluster initialization in the task of speaker diarization using agglomerative clustering. Speaker clustering is achieved by iteratively resegmenting the data into clusters and merging the most similar pair of clusters. The cluster initialization is the first step in this process, and it is very important as some clustering errors at this stage can not be corrected and generate poorer final results. The presented algorithm works in three steps. The first step finds likely speaker change points in the recording. The second step groups these segments (friends) together, creating the desired number of initial clusters (enemies between them). A third step ensures that all data is assigned to one of the clusters. We tested this algorithm on the RT05s dataset, obtaining an improvement of 13% relative DER and 2.5% absolute cluster purity.

6. Acknowledgements

We would like to acknowledge the Speaker Diarization group at ICSI for their thoughtful comments and Joe Frankel, Adam Janin and Jose Pardo for their help. This work was done during Xavier Angueras stay at ICSI within the Spanish visitors program.

7. References

[1] D. Reynolds and P. Torres-Carrasquillo, “Approaches and applications of audio diarization,” in *ICASSP’05*, Philadelphia, PA, March 2005, pp. 953–956.

[2] S. S. Chen and P. Gopalakrishnan, “Clustering via the bayesian information criterion with applications in speech recognition,” in *ICASSP’98*, vol. 2, Seattle, USA, 1998, pp. 645–648.

[3] J.-L. Gauvain, L. Lamel, and G. Adda, “Partitioning and transcription of broadcast news data,” in *ICSLP-98*, vol. 4, Sidney, Australia, 1998, pp. 1335–1338.

[4] C. Wooters, J. Fung, B. Peskin, and X. Anguera, “Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system,” in *Fall 2004 Rich Transcription Workshop (RT04)*, Palisades, NY, November 2004.

[5] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, “Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system,” in *RT05s Meetings Recognition Evaluation*, Edinburgh, Great Brittain, July 2005.

[6] J. Ajmera and C. Wooters, “A robust speaker clustering algorithm,” in *ASRU’03*, US Virgin Islands, USA, Dec. 2003.

[7] X. Anguera, C. Wooters, and J. Hernando, “Automatic cluster complexity and quantity selection: Towards robust speaker diarization,” in *Speaker Odyssey 06*, Puerto Rico, USA (to appear), June 2006.

[8] ——, “Speaker diarization for multi-party meetings using acoustic fusion,” in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Puerto Rico, USA, November 2005.

[9] X. Anguera, M. Aguilo, C. Wooters, C. Nadeu, and J. Hernando, “Hybrid speech/non-speech detector applied to speaker diarization of meetings,” in *Speaker Odyssey 06*, Puerto Rico, USA (to appear), June 2006.

[10] S. Shaobing Chen and P. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the bayesian information criterion,” in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, Feb. 1998.

[11] J. Ajmera, I. McCowan, and H. Bourlard, “Robust speaker change detection,” *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 649–651, 2004.

[12] N. Mirghafori and C. Wooters, “Nuts and flakes: A study of data characteristics in speaker diarization,” in *ICASSP’06*, Toulouse, France (to appear), May 2006.

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