The UVigo-GTM Speaker Diarization System for the Albayzin'10 Evaluation

Laura Docio-Fernandez, Paula Lopez-Otero, Carmen Garcia-Mateo

Department of Signal Theory and Communications, Universidade de Vigo
ldocio@gts.uvigo.es, plopez@gts.uvigo.es, carmen@gts.uvigo.es

Abstract

In this paper, the system submitted by the UVigo-GTM for the Albayzin 2010 Speaker Diarization Evaluation is described. This system is built upon the speaker segmentation system presented in our ICASSP 2010 paper. Specifically, the system uses a poisson-based false alarm reduction strategy. Then, the speaker segmentation strategy assumes that the occurrence of changes constitute a Poisson process, so changes will be discarded with a probability that follows a Poisson cumulative density function. The speaker clustering step we use an agglomerative clustering approach in which the speech segments are merged until reaching a stopping point.
Index Terms: speaker segmentation, speaker clustering, cluto

1. Introduction

Nowadays, an emerging application area where speech technologies are involved is the field of structuring the information of multimedia (audio-visual) documents. These multimedia documents are, in general, multi-speaker audio recordings, and for some applications it may be relevant to determine "who spoke when". This task is also referred to as "speaker segmentation and clustering" or "speaker diarization" in the literature. The segmentation of the data in terms of speakers could help in efficient navigation through audio documents, such as meeting recordings or broadcast news archives. Using these segmentation clues, an interested user can directly access a particular segment of the speech spoken by a particular speaker. Other applications of the speaker segmentation task include speaker adaptation in speech recognition and speaker identification-verification-tracking.

The Albayzin 2010 Speaker Diarization Evaluation task focuses in audio broadcast news programs. The UVigo-GTM speaker diarization system follows a two-stage speaker diarization approach: a speaker segmentation stage, which detects speaker change points; and a speaker clustering stage, where the speech segments, each spoken by one speaker, are clustered using an agglomerative hierarchical strategy.

In [1], an online four-step speaker segmentation system is introduced: it first performs a coarse segmentation of the data, then refines or discards the change points, discriminates between speech and non-speech, and merges segments that are likely to be spoken by the same speaker. It was noticed that this baseline segmentation system has a high false alarm rate and tends to estimate short segments. In [2], two novel approaches for reducing the number of false alarms, in order to avoid erroneous speaker changes, were introduced, evaluated and compared with the false-alarm discard algorithm proposed in [1]. The first approach rejects, given a discard probability, those changes that are suspicious of being false alarms because of their low ΔBIC value. The second strategy assumes that the occurrence of changes constitute a Poisson process, so changes will be discarded with a probability that follows a Poisson cumulative density function. The goal of such techniques is to confirm true speaker changes and suppress erroneous speaker changes. The UVigo-GTM speaker diarization system submitted for the Albayzin 2010 Evaluation is based on the second startegy for rejecting change-points.

To accomplish the clustering task, an agglomerative hierarchical clustering method was chosen. The clustering algorithm measures the similarity between clusters based on the similarity between pairs of speech segments. The critical elements of this clustering technique are the distance or similarity metric used to compare the speech segments, and the selection of the final number of clusters.

This paper is organized as follows. Section 2 gives a brief description of the baseline speaker segmentation system. The proposed approaches to reduce the false alarm rate are presented in Section 3. In Section 4 an explanation of the experimental framework is given. The performance of the speaker segmentation system using each one of the false alarm reduction strategies is shown and discussed in Section 5. Finally, Section 6 concludes this paper and gives some ideas of future work.

2. The speaker segmentation stage

The architecture of the baseline speaker segmentation system described in [1] is depicted in Fig. 1, where it can be observed that it has four stages: first, a coarse segmentation is made with the Distance Changing Trend Segmentation algorithm (DCTS) [3], in order to detect audio change-point candidates and then a refinement or rejection of these change-point candidates is performed by the Bayesian Information Criterion (BIC) algorithm [4]. After that, the system makes a Maximum a Posteriori (MAP) adaptation of three different Gaussian Mixture Models (GMMs) to decide whether the audio segment delimited by the new change-point and the preceding one is speech, music or silence/noise. If the segment is speech, the same procedure will be employed to classify the speech in male or female speech. Finally, when the two latest segments are speech, an approach based on the Cross Likelihood Ratio (CLR) [5] test is applied in order to find out if both speech segments are spoken by the same speaker; in that case both speech segments are merged.

2.1. Poisson distributed-based false alarm rejection strategy

The proposed strategy to discard false alarms is based on the monitoring of the ΔBIC value

$$\Delta BIC(i) = L(i) - \lambda P \quad (1)$$

where P is the penalty, corresponding to the number of free parameters of the Gaussian model, and λ is a weight that increases or decreases the influence of the penalty. When λ is a small

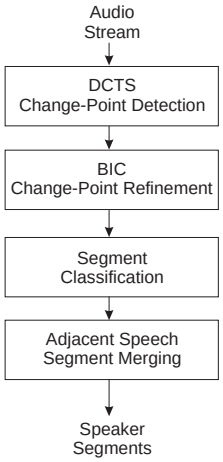

Figure 1: *Architecture of the speaker segmentation system presented in [1].*

value, less changes will be discarded by the BIC algorithm; the opposite happens when λ gets bigger.

Equation (1) has a member $L(i)$ which represents a likelihood ratio:

$$L(i) = \frac{L}{2}log|\Sigma| - \frac{L_1}{2}log|\Sigma_1| - \frac{L_2}{2}log|\Sigma_2| \qquad (2)$$

where L, L_1 and L_2 are the number of frames of segments X, X_1 and X_2 respectively; and Σ, Σ_1 and Σ_2 are the covariance matrices of the models M, M_1 and M_2 respectively. Thus, there will be a change in the audio stream when

$$\frac{L}{2}log|\Sigma| - \frac{L_1}{2}log|\Sigma_1| - \frac{L_2}{2}log|\Sigma_2| > \lambda P \qquad (3)$$

In this false alarm suppression strategy, it is assume that the occurrence times of change-points can be modeled by a Poisson process.
A Poisson process is an independent occurrence process where the number of occurrences in two disjoint time intervals is independent, the probability of having an occurrence is proportional to the observed interval, and the occurrences are not simultaneous.

The process we are dealing with in speaker segmentation fulfills four properties, as it is a process where arrivals (of changes) happen independently from the others and in random instants. Poisson processes have a probability density function

$$f(\mu, x) = \frac{e^{-\mu}\mu^x}{x!} \qquad (4)$$

and its cumulative density function (cdf) is the sum of the probability density function in all the points below a given value:

$$F(\mu, x) = \sum_{i=0}^{x} \frac{e^{-\mu}\mu^x}{x!} \qquad (5)$$

The parameter μ represents the mean of the distribution. In this case, it will represent the number of expected changes.
The properties of the Poisson distribution are going to be used as follows: μ occurrences are expected in a given period of time. Therefore, initially a change will be accepted with a very high probability, but as the number of accepted changes increases and gets close or over the expected number, they will be more likely to be rejected. This is easily modeled by using the cumulative density function $F(\mu, x)$ as a discard probability: this discard probability will be very low at first, and as the mean is approach or exceded, it will get bigger and bigger, until a moment where it will be close to 1 (this means that all the occurrences will be rejected). It can be seen in figure 2 how the discard probability increases as the number of accepted changes gets bigger.

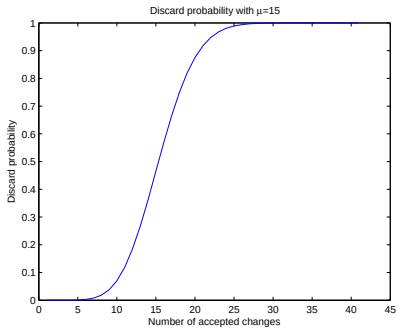

Figure 2: Discard probability on the Poisson-based discard algorithm.

3. The speaker clustering stage

The speaker clustering stage is based on an agglomerative hierarchical clustering technique and standard speaker recognition techniques based on GMM-UBM models are used.

In a training stage, a UBM-GMM model Θ, is constructed using the audio training data. In the clustering stage, first a segment model θ_i is derived by MAP adapting the Θ model parameters using the acoustic frames of the speech segment x_i. Next, a hierarchical classification algorithm is applied in three steps: 1) the first step consists in computing some distance or similarity measure between each pair of speech segments; 2) the second step consists in creating a tree structure by starting with each segment in its own cluster, and recursively merging clusters according to some distance-related criterion; 3) and the last step is to choose one of the partitions, a process called tree cutting.

Several measures of similarity between segments or clusters can be used. The first experiments were conducted using a distance measure which uses information about the likelihood score between pairs of models and speech segments. Specifically, each speech segment, x_i, is scored against all the set of trained segment models, θ_j, and the collection of those scores is used to form a N-dimensional vector X_i (N speech segments) that represents the speech segment x_i in a multidimensional vector-space:

$$X_i = \{ll(x_i|\theta_1), ll(x_i|\theta_2), \cdots, ll(x_i|\theta_N)\}$$

$ll(x_i|\theta_j)$ is the log-likelihood of the ith speech segment x_i given the model of the jth speech segment model θ_j. Thus, the similarity between two speech segments can be computed straightforwardly by using the cosine distance between the two corresponding vectors. This distance ignores absolute sizes of the measurements, and only considers their relative ones; and it is a popular distance measure for comparing documents in the information retrieval literature.

Hierarchical agglomerative methods are well documented in the literature. The aim is to pick the closest pair of clusters according to a distance matrix and merge them. This step is repeated until there is only one cluster. The distance matrix only gives the distance between pairs of single data points, so some method is required to construct a distance between clusters from distances between single data points. There are several possibilities, being most of them variants of the single-link, complete-link, group average-link and minimum variance algorithms. Among these algorithms, the single-link, complete-link and group average-link are the most popular. These algorithms differ in the way they characterize the similarity between a pair of clusters. In the single-link method, the distance between two clusters is the minimum of the distances between all pairs of patterns drawn from the two clusters. In the complete-link algorithm, the distance between two clusters is the maximum of all pairwise distances between patterns in the two clusters. In the group average-link approach, the distance between two clusters is the average of all pairwise distances between patterns in the two clusters. Very little is known about what qualities make a cluster distance good for clustering. The general purpose clustering toolkit, CLUTO, developed by the University of Minnesota [7], was used for this unsupervised speaker clustering stage.

In the results submitted to Albayzin 2010 Speaker Diarization Evaluation, the complete-link algorithm was selected, and the stopping criterion was based on a fixed number of clusters, specifically the number of clusters was fixed to 90 for each audio file.

4. Experimental framework

4.1. Database

The training and evaluation database consists of Catalan broadcast news data from the 3/24 TV channel that was recorded by the TALP Research Center from the UPC, and was annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project, funded by the Generalitat de Catalunya. The Corporaci Catalana de Mitjans Audiovisuals, owner of the multimedia content, allows its use for technology research and development. The database, that includes around 87 hours of sound (24 files of approximately 4 hours long), was splitted into two parts: one part for training/development (2/3 of the total amount of data), and the other part for evaluation (the remaining 1/3).

The number of speakers per recording ranges from 30 up to 250. This high number of speakers is due to the domain of the data. Some speakers are common among different recordings. That is the case of the newscaster, the journalists or some voices from the commercials, etc. However, most of the speakers have short duration turns since their presence depends on the news itself.

The 16 available files to perform the training/development of the segmentation system were splitted as follows:

- Sessions 1 to 8 and 10 to 15: training of the silence, speech, music and GMM-UBM models.
- Sessions 9 and 16: selection of the parameters that achieve the best performance. The parameters to select were μ, λ, M.

After testing on the development data the selected parameters were: $\mu = 15.0$, $\lambda = 2.5$, $M = 64$.

4.2. Metric

Diarization Error Rate (DER) as defined by NIST in Rich Transcription evaluations [8] will be used to assess the submitted systems. In order to measure the performance, an optimum one-to-one mapping of reference speaker IDs to system output speaker IDs is computed. The measure of optimality will be the aggregation, over all reference speakers, of the time that is jointly attributed to both the reference speaker and the (corresponding) system output speaker to which that reference speaker is mapped. This mapping over will always be computed over all the speech, including regions of overlap . Mapping is subject to the following restrictions:

- Each reference speaker will map to at most one system output speaker, and each system output speaker will map to at most one reference speaker.
- Mapping of speakers will be computed separately for each speech data file.

Speaker detection performance will be expressed in terms of the miss and false alarm rates that result from the mapping. An overall time-based speaker diarization error score will be computed as the fraction of speaker time that is not attributed correctly to a speaker.

4.3. Acoustic features

The audio signal is characterized by 12 mel-frequency cepstral coefficients (MFCC) extracted every 10 ms using 25 ms Hamming windows. Then these cepstral features are augmented by the log-energy. The DCTS and BIC change detection stages use only the 12 MFCCs and the log-energy as features. In the speech/non-speech classification and the gender classification modules the first and second derivatives of this feature vector are also considered.

The speech, non-speech, male and female and GMM-UBM models are 64 diagonal Gaussian Mixture Models (GMM) directly trained on data extracted from the train corpus by using the Expectation-Maximization (EM) algorithm.

5. Experimental results

Table 1 provides the results obtained by the system submitted by the UVigo-GTM research group. In column 2 of Table 1 it can be observed a lack of consistency in the speaker diarization error between different speech (session) files. This fact suggest that a serious mistake was made somewhere when processing these evaluation files with the speaker diarization system. After checking the code and the file processing the mistake was corrected and the results obtained are those shown in column 3 of Table 1. These results are still far from those obtained by the other participants.

A manual inspection of the number of speakers in each evaluation audio file shows that selecting 90 as the number of clusters is not the best option for all the audio files. The wrong choice of number of clusters affects the system performance adversely. Table 2 shows the number of speakers in each audio file. Selecting the right number of clusters has not been considered in the submitted system. The approach used a fixed number of clusters for all the audio files.

Speaker Diarization experiments were conducted using the "Unweighted Pair-Groups Method Average (UPGMA)" criterion function for agglomerative clustering, which defines cluster similarity in terms of the average pairwise similarity between the segments in the two clusters. This criterion is widely used

Table 1: *Speaker diarization results on the evaluation corpus: CLINK agglomerative criterion function and 90 clusters per audio file.*

audio file	DER	
session17	40.49	40.49
session18	68.33	42.99
session19	40.11	40.11
session20	69.70	45.36
session21	42.80	42.80
session22	88.23	39.61
session23	70.57	34.87
session24	36.96	36.96
global	58.03	40.21

Table 2: *Number of speakers in each evaluation audio file.*

s17	s18	s19	s20	s21	s22	s23	s24
106	91	70	120	65	93	66	93

in text document clustering because it is more robust than other traditional agglomerative clustering approaches. The resulting SDER are shown in Table 3. Compared with the results on Table 1, the use of UPGMA gives significant improvement over the submitted results. Table 3 also shows the influence played by the number of clusters in the speaker diarization error.

Table 3: *Speaker diarization results on the evaluation corpus: UPGMA agglomerative criterion function.*

	DER						
	70	80	90	95	100	105	110
session17	43.44	42.54	37.98	37.63	37.61	**35.43**	37.20
session18	38.46	36.67	35.69	34.41	34.66	34.21	**34.01**
session19	**29.87**	32.57	32.84	32.22	32.89	33.15	33.95
session20	38.58	38.12	38.73	37.69	37.60	36.94	**34.68**
session21	**32.01**	32.59	31.91	32.09	32.95	33.73	35.27
session22	44.91	40.98	39.68	39.18	38.80	38.26	**38.07**
session23	**27.58**	28.89	28.71	28.57	28.62	28.83	29.30
session24	32.66	33.60	34.32	34.20	**32.75**	33.29	32.88
global	36.09	35.87	35.11	34.60	34.54	34.23	34.34

6. Conclusions and future directions

The speaker segmentation system submitted to Albayzin 2010 Evaluation was described in this paper. The speaker diarization task focuses in the context of broadcast news. According to the results obtained by the proposed system on the evaluation dat, it was realized that a huge mistake was made when processing some evaluation speech files.

Future work will focus on combining the traditional short-term MFCCs features with prosodic and other acoustic features in order to discriminate better between speakers. Related to the speaker clustering stage, future work will focus on:

- The analysis of methods for speaker clustering used in the state-of-art speaker diarization systems.
- The analysis of strategies to detect or to discover the number of clusters, i.e, approaches for cluster stopping.
- The use of other similarity measures between speech segments and other criteria to group speech segments.

7. Acknowledgements

The UVigo-GTM group would like to thank the organizers and coordinators of the Albayzin 2010 Speaker Diarization Evaluation for their help and for kindly providing the data corpus. This work was partially sponsored by the Spanish Ministry of Science and Innovation under the project TEC2009-14094-C04-04.

8. References

[1] L. Docio-Fernandez, P. Lopez-Otero, and C. Garcia-Mateo, "An adaptive threshold computation for unsupervised speaker segmentation," *Proc. 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, September 2009, pp. 849–843.

[2] P. Lopez-Otero, L. Docio-Fernandez and C. Garcia-Mateo, "Novel strategies for reducing the false alarm rate in a speaker segmentation system," *Proc. 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2010*, Dallas, USA, March 2010, pp. 1470–1473.

[3] Y. Wang, J. Han, H. Li, and T. Zheng, "A novel audio segmentation method based on changing trend of distance between audio scenes," *Journal of Communication and Computer*, vol. 3, pp. 22–30, July 2006.

[4] Scott S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, February 1998.

[5] D. A. Reynolds, E. Singer, B. A. Carlson, and J. J. McLaughlin, "Blind clustering of speech utterances based on speaker and language characteristics," *in Proc. of the Int. Conf. on Spoken Language Processing*, Sydney, Australia, 1998.

[6] A. O. Allen, "Probability, Statistics, and Queueing Theory with Computer Science Applications", 2nd ed. Academic Press, 1990.

[7] "CLUTO - Software for Clustering High-Dimensional Datasets", http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview.

[8] "NIST Rich Transcription Evaluation", http://nist.gov/itl/iad/mig/rt.cfm.