Blind Segmentation and Labeling of Speakers via the Bayesian Information Criterion for Video-Conference Indexing

Javier Ferreiros, Javier Macías-Guarasa, Juan-Manuel Montero, Rubén San-Segundo, Luis F. D'Haro and Francisco-Javier Yuste

Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica
Universidad Politécnica de Madrid

{jfl, macias, juancho, lapiz, lfdharo, jyuste}@die.upm.es

Abstract

The purpose of this paper is to present a system which breaks input speech into segments and identifies each new appearance of the same speaker with a consistent label. This task adds up to a topic detection system that makes use of key-word recognition to obtain suitable labels for an automatic indexing system project. Both the segments definition and the identification of the speaker for each segment are performed using an acoustic similarity measure.

Our task is to separate and identify the different speakers who appear in a video-conference session without any prior knowledge of the speakers or their number. The first aim is to detect the time points where a speaker change takes place using a robust acoustic change detection (ACD) system. Afterwards, the regions defined by these time marks are labeled with the use of a clustering algorithm. The Bayesian Information Criterion (BIC) is the key element in the system, and is used in several ways as a measure to compare speech. EERs of 13.66% are obtained for this task with a soft feeding back of clustering information to enhance ACD performance.

1. Introduction

Under the contract TIC2000-0198-P4-04 ISAIAS with the Comisión Interministerial de Ciencia y Tecnología, a project for automatic indexing of video-conference sessions have been developed. The aim is to end up with a system that prepares the proper labels as to be able to answer questions as "Play for me what [username] told about [topic]".

In this paper we discuss the part of the project concerning the blind speaker recognition needed. Our database consists of a set of audio files of three-hour length containing different video-conference sessions, including speech from both genders of medium age speakers. The speech comes from an unknown and widely varying number of speakers.

In the system, we first employ an ACD procedure on the input speech resulting in a set of acoustically homogeneous segments. This basic ACD is enhanced via the clustering of these segments. Finally, an additional clustering stage and a filter for silence segments is included to obtain the definitive labels.

Although our system has been designed for indexing video-conference sessions, the same system has been also tested with other recordings from radio broadcast news with similar performance.

2. System Architecture

Our system is composed of the modules shown in Fig. 1. First of all, we extract the features of the signal: 10 unsmoothed cepstral coefficients plus the frame energy for each 10 ms. frame. That number of features has proved to be enough to achieve acceptable performance using full-covariance Gaussian models [1]. The following module generates a set of initial hypothesis for the ACD. The purpose is to save computing power not hypothesizing a break point for each single frame. A hypothesis can be defined as the time point (frame) in which an acoustic change might be present.

The Acoustic Change Detector, ACD, is the module which decides whether a hypothesis will be validated as an acoustic change or not.

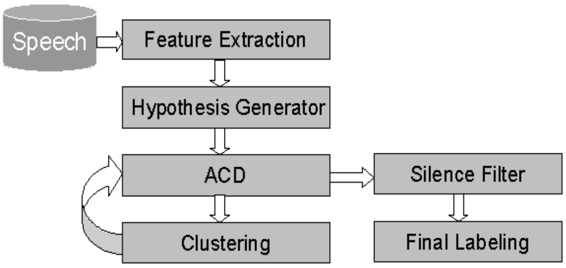

Figure 1: System Architecture

This decision is taken using the BIC. After this, a clustering of the segments with similar acoustics is carried out using the same criterion than in the ACD.

The information conveyed in this clustering of segments is fed back to the ACD to refine the decisions making them more robust as it will be shown below. A simple filtering of this segments is implemented to eliminate those with non-speech events. Finally, a new clustering of the final segments identifies them with a proper speaker label.

2.1 Break-point hypothesis generator

Once we have the speech features for each frame, instead of hypothesizing a break point in every frame of the input signal, we select a reduced set of frames for which the system estimates that the likelihood of an acoustic change is high enough. This set will contain the most silent regions in the recording. This allows us to save an important number of calculations in the ACD procedure. The hypotheses of the ACD are the initial and final time points of the hypothesized segments. To decide whether a frame belongs to a silence region or not we only take into account the frame energy.

It is essential to know the characteristics of the silence and the speech with respect to their energy and automatically obtain them from the new recordings to be processed. As a first step, we obtain the energy histogram and set a threshold to split it into two regions. An example is shown in Fig. 2, where we represent the energy in dB in abscissas plotted against the number of frames.

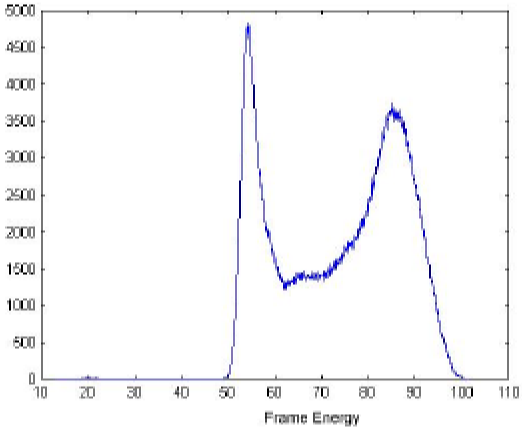

Figure 2: Example of an energy histogram.

The system estimates a first threshold separating silence from speech as the minimum smoothed-histogram value between both maximums. Under this initial threshold, we calculate a Gaussian model for the noise, while the rest of the frames are used to generate the speech model. The maximum likelihood estimation can be used to select the adequate class for every frame:

$$\frac{x - m_n^2}{2 \cdot s_n^2} + \ln\left(\frac{1}{s_n}\right) < \frac{x - m_s^2}{2 \cdot s_s^2} + \ln\left(\frac{1}{s_s}\right) \qquad (1)$$

where μ_n, μ_s and σ_n, σ_s are the means and standard deviations of noise and speech, respectively. We have taken logarithms in both terms for easier computing.

Because of the difference between the variances of the two models (speech has always much more energy variance than noise), the speech Gaussian overcomes the noise Gaussian in the region of low energies. The problem can be seen in Fig. 3, where we observe that the expression 1 would classify the frames whose energy is lower than the lowest energy cross point of both curves (~40 dB.) as speech instead of noise.

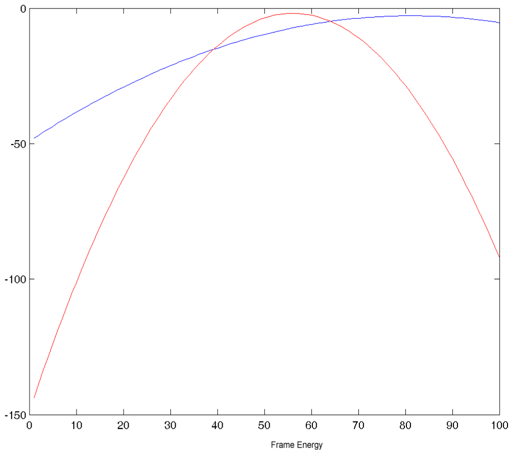

Figure 3: Log-likelihood vs. frame energy for speech (wider curve) and background noise.

To avoid these errors a simple solution is proposed: We choose the highest energy cross point as the final threshold (~63 dB. in Fig. 3). The comparison of the frame energies with this threshold gives us the classification criterion: if the frame energy is higher than the threshold, the frame is labeled as speech, otherwise the frame is considered noise.

A median filter will smooth up the result for consecutive frames to obtain the final hypothesized break points.

2.2 The Acoustic Change Detection System

Our approach is based on the ideas shown in [1]. Here, we present the basis of the Acoustic Change Detection system that can be also found in [2] and [3], underlining the differences and particularities of our new system.

The Bayesian Information Criterion (BIC), a likelihood measurement penalized by the complexity of the assumed model, is used as the model selection criterion because of its properties: robustness, threshold dependence desensitization and optimality. This measure will be used widely in both the ACD and in the clustering stages. For adjacent acoustic segments (delimited by break points that are hypothesized by the previous module), an actual break point is inserted by comparing the fit of a single multidimensional Gaussian model for the entire segment to the conjunct fit of separate models for each side of the break. We compare these two alternatives using the BIC. Given a set of N vectors X = { x_i : i =0 … N-1 }, that we want to represent by a model M, the BIC is calculated as follows:

$$BIC(M) = \log\left[L(X,M)\right] - \frac{l}{2}\cdot\#(M)\cdot\log(N) \qquad (2)$$

where the penalty weight, λ, should be 1, at least in theory[1]. #(M) is the complexity of the model measured by its free parameter count and L(X,M) is the likelihood of data X under model M.

For each hypothesized break point, we have to decide if the whole segment comes from the same acoustic conditions or, in contrast, there are two acoustically different segments that we should break apart using the break point. Using Gaussian models we have:

H_0: $x_0 \ldots x_{N-1} \sim N(m, \Sigma)$

H_1: $x_0 \ldots x_{N_1-1} \sim N(m_1, \Sigma_1)$

$x_{N_1} \ldots x_{N-1} \sim N(m_2, \Sigma_2)$

where N_1 represents the hypothesis break point. We arrive to the following hypothesis test, assuming that $N_2 = N - N_1$:

$$\log\left(\frac{|\Sigma|^{N}}{|\Sigma_1|^{N_1}\cdot|\Sigma_2|^{N_2}}\right) - \frac{l}{2}\cdot\left(d + \frac{d\cdot(d+1)}{2}\right)\cdot\log(N) \underset{H_0}{\overset{H_1}{\gtrless}} 0 \qquad (3)$$

If the equation 3 is positive, we decide H_1 and break the whole segment into the two sub-segments. In this expression, the complexity of the model is penalized via the factor d + d(d+1)/2, i.e. the number of free parameters in a full covariance Gaussian model for d-dimensional feature vectors.

In our system, unlike the procedure described in [1], where the non-speech regions were not processed by the ACD, we process the full input signal. The reasons are mainly two: we have a much less powerful and reliable hypothesis generator than in [1] and, moreover, in the database used in [1] silence regions were not common, while our video-conference task frequently contains large segments without speech.

2.3 Clustering to enhance ACD

The clustering of the segments generated in the ACD can be used to improve the ACD performance feeding back its information. The same acoustic measure, the BIC, is used in the clustering algorithm. As a result of this clustering, we obtain a set of clusters, agglomerating a set of segments with homogeneous acoustic conditions.

The basic iteration of the clustering algorithm is as follows:

1) Pick one ACD segment.
2) Remove this segment from its cluster and update cluster data.
3) Find the "closest" cluster (if any) to the segment.
4) If there is a representative cluster, go to 6)
5) Generate a new cluster with only this segment in it. Goto 7)
6) Update this cluster with the segment information.
7) If last segment then stop, else go to 1)

We perform 10 iterations like this one (or come to an end when not enough average distance improvements are obtained). We use likelihood values as distances between data sets and define a distance between two acoustic segments as:

$$d(X,Y) = (N_X + N_Y)\cdot\log|\Sigma_{X\cup Y}| - N_X\cdot\log|\Sigma_X| - N_Y\cdot\log|\Sigma_Y| \quad (3)$$

where X and Y are two data sets that can be both single acoustic segments (we will use the letter S in these occasions) or clusters of agglomerated segments (for which we will use the notation C). The idea is to consider the distance between the representative clusters for both segments to be compared:

$$G = d(C_1, C_2) \qquad (4)$$

where C_i is the cluster to which the segment S_i belongs (is nearer to).

These clusters are supposed to represent well the segments to compare. Thus, the distance between these representative clusters should be relevant to make the ACD more robust. The breaking decision will not be taken only on direct comparison of the two segments that some times might be too short, but will also use the distance between the representative clusters of the two segments. Because the clusters agglomerate all the segments in the file with similar acoustics, much more robust information is now used to make the decision.

There are different options in the integration of the clustering into the ACD. A summary of the strategies we have tested can be seen in table 1. In the “hard

[1] We have found throughout our experiments that the value of *l* had to be tuned in order to end up near the EER.

decision" alternative, we only allow a break between segments coming from different clusters. The second "soft integration" alternative calculates a linear combination of the distances between segments and between clusters to obtain the actual distance. It has a parameter α that has to be tuned on some training material.

Strategy	Validation of hypothesis
Hard decision	$d(C_1,C_2)\neq 0$
Soft integration	$(1-a)\cdot d(S_1,S_2)+\frac{s_{d(S_1,S_2)}}{s_{d(C_1,C_2)}}\cdot a\cdot d(C_1,C_2)$

Table 1: Summary of the strategies for clustering integration into ACD

2.4 Silence Filter

In this video-conferencing task, a filtering of the silence chunks must be done following the reasons given above. The filtering rate, defined as the percentage of frames labeled as non-speech to decide that a segment does not contain speech, has been tuned so that the chunks with some speech inside will not be filtered out.

2.5 Final Labeling

A new clustering of the remaining speech segments is performed. The clustering system is similar to the one described in Section 2.3. λ has to be set considering a tradeoff between two behaviors of the system:

- If we use a low λ, the system may detect slight changes of acoustic conditions even within the same speaker speech segments.
- If we use a high λ, the system may assign the same label to utterances coming from different speakers.

λ optimization is performed on a hand labeled training database.

3 Experiments and results

We have tested our system on a 45 minutes hand-labeled video-conference session including speech from 11 speakers. Following the evaluation directions in [1] and [2] for the ACD task, we have tested both options in the feeding back of the clustering information to the ACD with the results shown in table 2. Our best result has been achieved by the soft integration of the clustering information.

Strategy	EER
Baseline ACD	19.42 %
Hard feedback	15.70 %
Soft integration	13.66 %

Table 2: Summary of the best results

4 Conclusions

In this paper, we present a system mainly based in [1] which performs the speaker change detection and identification. We have used a simpler hypothesis generator based on the comparison of the frame energies with an estimated threshold to classify the frames into the noise and speech classes. With respect to the ACD, a soft integration has been found to be the best strategy for feeding back the clustering information into the ACD module. The result is a robust, accurate and reliable system that matches the purposes of the video-conference indexing system for which it has been designed, performing 13.66% EER.

5 Acknowledgements

This work has been done under the contract TIC2000-0198-P4-04 ISAIAS with the Comisión Interministerial de Ciencia y Tecnología to whom we want to thank for their support. We also want to thank the rest of the Speech Technology Group for their collaboration and support.

6 References

[1] Ferreiros López, Javier and Ellis, Daniel P. W., "Using acoustic condition clustering to improve acoustic change detection on broadcast news", *ISCLP 2000*.

[2] Daben Liu, Francis Kubala, "Fast speaker change detection for broadcast news transcription and indexing", *Eurospeech 99*.

[3] Scott Shaobing Chen y P.S. Gopalakrishnan. "Speaker, environment and channel change detection and clustering via the bayesian information criterion". DARPA, 1998. Broadcast News Transcription And Understanding Workshop.