Broadcast News Speaker Tracking for ESTER 2005 Campaign

Dan Istrate, Nicolas Scheffer, Corinne Fredouille and Jean-François Bonastre

Laboratoire d'Informatique d'Avignon (LIA)
339, chemin des Meinajaries; Agroparc BP 1228
84911 Avignon Cedex 9; France
(Nicolas.Scheffer,Dan.Istrate,Corinne.Fredouille,jfb)@lia.univ-avignon.fr

Abstract

This paper presents the speaker tracking system of the LIA laboratory, validated during ESTER 2005 campaign on a radio broadcast news corpus of about 90 h. The LIA speaker tracking system firstly uses an acoustic class segmentation in order to suppress non speech frames and to detect the speech conditions. Secondly, a speaker diarization process is applied in order to provide speaker detection system (the last step) with speaker homogeneous segments (boundaries and clustering). The speaker detection system uses UBM/GMM likelihood ratios in order to decide if a segment belongs to one tracked speaker. The speaker tracking system is presented and some results obtained during ESTER 2005 campaign are proposed. The presented systems are based on the ALIZE platform (Automatic speaker recognition C++ library).

1. Introduction

For the past decades, the amount of multimedia data has increased hugely, involving large difficulties to manage it. Consequently, information extraction in order to index multimedia recordings is an active research domain. Different types of information may be extracted from sound signals: sound event, speaker information (turns, identity, gender), transcription, linguistic information, etc; each being related to a specific task. For example in broadcast news, a request may be to find the speech segments pronounced by a known speaker, referred to as the speaker tracking task.

Indeed, the speaker tracking task consists in detecting portions of the document that have been uttered by a given speaker known beforehand and for which training data are available before the test/tracking stage.

This paper presents the speaker tracking experiments done by the LIA during ESTER 2005 campaign [1, 2]. The LIA speaker tracking system follows a three steps algorithm (illustrated in Figure 1). Firstly, an acoustic segmentation is applied, in order to remove non-speech parts of signal as well as for detecting the speaker gender and recording conditions. Secondly, a speaker turn detection is done. This second step is necessary to provide speaker detection process with speaker homogeneous segments. It is done by a speaker diarization system which carries out speaker turns detection as well as speaker clustering. Finally, a speaker recognition system is applied, on each separate segment or on the different clusters (all the segments of a cluster are supposed to belong to one speaker only).

In this paper, the acoustic segmentation and the speaker diarization are firstly evaluated separately in order to show their intrinsic performance. Two different strategies are then proposed to link the speaker diarization to speaker detection process:

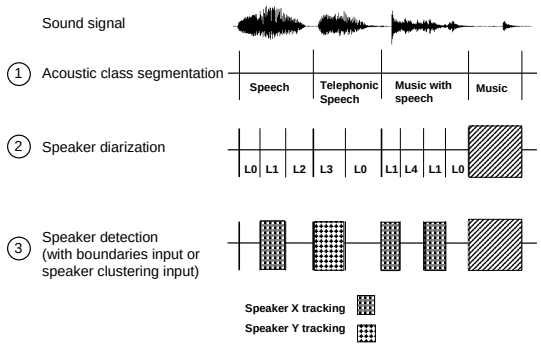

Figure 1: Speaker tracking system

- the only use of the speaker boundaries provided by speaker diarization
- the use of all the speaker information provided by speaker diarization (boundaries and clustering)

Both the link strategies are evaluated in the framework of the ESTER 2005 campaign.

The software has been developed with ALIZE toolkit [3, 4], and is available thanks to an open software licence.

Section 2 presents the ESTER evaluation, which constitutes the experimental basis of this work. Section 3 is dedicated to the acoustic segmentation, section 4 to the speaker diarization system and section 5 to the speaker detection part of the whole system. The speaker tracking system as well as the two strategies are evaluated in section 6. Finally, section 7 concludes this paper.

2. ESTER 2005 campaign

The presented speaker tracking system has been validated during the ESTER 2005 campaign. This campaign has focused on the evaluation of rich transcription and indexing of radio broadcast news in French. It implements three categories of tasks, namely transcription (T), segmentation (S), and information extraction (E). In the segmentation category, different tasks were proposed like: sound event tracking (acoustic segmentation), speaker diarization and speaker tracking.

The corpus of ESTER 2005 campaign is about 90h of manually transcribed radio broadcast news for training purposes, a part of 8h was identified as a development set, named *Dev*. This acoustic corpus contains shows from four different sources, namely France Inter, France Info, Radio France International (RFI) and Radio Télévision Marocaine (RTM). The test set, named *Eva*, consists in 10 hours of radio broadcast news taken

from the four stations of the training corpus, France Culture for which only non transcribed data were available, and Radio Classique for which no specific training data was available.

The performance was evaluated in terms of F-measure, defined as $2RP/(R+P)$ where:

$$R = \frac{\sum_i t(c_i;c_i)}{\sum_i t(c_i;c_i) + t(\overline{c}_i;c_i)} \text{ and } P = \frac{\sum_i t(c_i;c_i)}{\sum_i t(c_i;c_i) + t(c_i;\overline{c}_i)}$$

where R is the recall and P is the precision.

As the interpretation of the F-measure is not intuitive, errors are also analyzed in terms of miss and insertion rates. Errors are computed based in time marks, in seconds, with a tolerance of 0.25s for the reference segment boundaries.

For the diarization task, a specific performance measure [5] is considered in order to take into account deletions and insertions of speech in addition to speaker substitutions, after optimal matching between true and arbitrary speaker names.

3. Acoustic class segmentation

Acoustic class segmentation is the first step of the proposed system and aims at detecting speech/non speech segments, acoustic conditions and speaker gender. The acoustic recording conditions and the speaker gender can improve tracking system performance. For example, the use of specific models for male/female or for telephone/studio conditions is useful for the speaker detection (section 5.4).

The ESTER sound event tracking task consists in identifying, on the one hand, parts of the document containing music, whether in the foreground or in the background, and, on the other hand, parts of the document containing speech, possibly with background music.

The acoustic macro-class segmentation is composed of two step segmentation (hierarchical segmentation). During the first step, the signal (a radio show) is segmented in: speech (S), telephonic speech (TS), music (M) and music with speech (MS). The algorithm is based on an ergodic HMM with four states and on a Viterbi decoding procedure.

The gender detection is the second step. Each "speech" segment (S/MS[1],TS) is segmented and labelled as male or female. This step uses also a HMM based method, with 4 states: speech male, speech female, telephonic speech male and female. Each state has a gender and acoustic dependent model. More details may be found in [6].

Table 1: Sound event tracking error rate obtained on the *Eva* corpus in terms of false alarms, false rejections and F-Measure

Event	fa	fr	F-Measure
Music	10.9 %	38.7 %	0.55
Speech	36.6 %	0.7 %	0.99

The results of sound event tracking task on the *Eva* corpus for music and speech tracking are presented in table 1. The high false alarms rate for speech segments is due to large disproportion of overall duration between speech and music events (the test set contains less than 5% of non speech segments - music for example). No specific work was done in music detection explaining relatively poor results.

[1]speech and music with speech segments are merged into a single speech class

4. Speaker diarization

Speaker diarization (SRL) aims at segmenting documents into speaker turns and at grouping together portions of the document uttered by the same speaker. Speakers are not known beforehand and identification is not required. This task is usually implemented in two steps: speaker boundary detection and speaker clustering (same speaker segment grouping).

The LIA speaker diarization system uses a one step algorithm based on E-HMM (Evolutive HMM) [6, 7, 8]. Each E-HMM state characterizes a particular speaker and the transitions represent the speaker changes. All possible changes between speakers are authorized.

During the segmentation and clustering process, a HMM is generated with a state for each detected speaker. This algorithm has 2 stages: segmentation and resegmentation. The segmentation stage is launched separately on each acoustic macro-class (SH, SF, TSH and TSF). It is composed of 4 steps:

1. **Initialisation**: A first model, named L_0, is estimated on all speech data to process. The HMM has one state, L_0 state.
2. **New speaker detection**: A new speaker is detected in the segments labelled L_0 as follows: a segment is selected among all the L_0 segments by likelihood maximization criterion. This selected segment is then used to estimate the model of the new speaker, named L_x, which is added to the HMM.
3. **Adaptation/Decoding loop**: The objective is to detect all segments belonging to the new speaker L_x. All speaker models are re-estimated through an adaptation process according to the actual segmentation. A Viterbi decoding pass is done in order to obtain a new segmentation. This loop adaptation/decoding is re-iterated while the segmentation is not stable.
4. **Speaker model validation and stop criterion**: The current segmentation is analyzed in order to decide if the new added speaker, L_x, is relevant. In this case the decision is made according to heuristical rules on speaker L_x segment duration. The stop criterion is reached if there is no more segment available in L_0. On the contrary, the process goes to the step 2.

The resegmentation stage is launched on all the recordings, after merging the four segmentations (obtained after applying the segmentation process on each acoustic class). This stage aims at correcting, if possible, the acoustic macro-class segmentation errors (a speaker divided into speech and telephonic speech), at refining the boundaries and at deleting unreliable speakers.

This stage is based only on the third step of the segmentation process. A HMM is generated from the segmentation and the adaptation/decoding loop is launched. At the end of each iteration, speakers with too short duration are deleted.

4.1. HMM transition probability calculus

All transition probabilities from state i to state j ($i \neq j$) are equiprobable. The ratio between the intra-state probabilities P_{ii} and inter-state probabilities P_{ij} (with $i \neq j$) remains constant. These probabilities are calculated according to equation

1, where γ is fixed to 9.

$$P_{ii} = \frac{\gamma}{\gamma + N - 1}, \quad P_{ij} = \frac{1}{\gamma + N - 1} \tag{1}$$

Concerning the Viterbi decoding step, a multiplicative coefficient (fudge) f is applied to transition probabilities in order to make them comparable with likelihoods (f is fixed to 2).

4.2. Results

The results of speaker diarization system are presented in Table 2. The used acoustical parameters are 20 LFCC (estimated every 10 ms on the frame of 20 ms) and logarithmic energy. No normalization is applied on the acoustical parameters.

Table 2: Speaker diarization performance presented in terms of miss speech error (*Mi*), false detected speech error (*Fa*), speaker segmentation error (*Spk*) and global error (*All*) on the *Dev* and *Eva* corpus

Dev				*Eva*			
Mi	Fa	Spk	All	Mi	Fa	Spk	All
1.5	0.1	15.3	16.9	0.6	0.0	17.4	18.0

We can observe that the results on the *Dev* corpus are comparable with results on the *Eva* corpus. Regarding the performance on individual shows (not provided in this paper), it can be observed that the global segmentation rate presents a large variability (from 0% to 44% of global error). Underestimate of the speaker number has been also observed (242 instead of 361 on the whole *Eva* set).

5. Speaker tracking

Speaker tracking is somewhat similar to sound event tracking when speakers are the events to track. This task consists in detecting portions of the document that have been uttered by a given speaker known beforehand and for which training data is available before the test stage.

The speaker detection part of the tracking system relies on LIA speaker detection system [4]. It has two stages: world and speaker training models and speaker detection. During the training stage, the reference speaker boundaries were used and the acoustic macro-class segmentation system (described in section 3) was applied on the corresponding segments (since the acoustic conditions were not available in the references).

5.1. Signal parametrization

The acoustical parameters are 16MFCC and its first derivatives. They are normalized depending on the gender and acoustic conditions, which implies that the mean and the variance of each record become 0 and 1, respectively.

5.2. World model training

The world model is trained using EM/ML algorithm. Different world models are estimated: gender dependent (male/female) and acoustic condition dependent (telephone and studio). Only a quarter of data of each radio has been used in order to reduce computation time. The speaker specific training data have been included in the world model training phase. The world models are composed of 1024 Gaussian components with diagonal covariances.

5.3. Speaker model training

The speaker models are adapted from the world model, using a mean only EM/MAP technique [9] on multi acoustic condition data. The total amount of data has been limited to 20 minutes and the reference speaker diarization is used.

5.4. Speaker detection

A likelihood ratio computation (eq. 2) is performed in order to detect the speakers using only the 10 best Gaussians selected by the world model on each frame.

$$LLR = \frac{l(y|X)}{l(y|W)} \tag{2}$$

where y is the analyzed frame, X the client model and W the world model. The final LLR is the geometric mean of the frame LLR.

5.5. Score normalization

A T-Norm score normalization is applied [10]. T-Norm consists in computing the mean and variance parameters of the scores between the current test data (the segment in interest) and a set of impostor models. Then the score is normalized following the formula:

$$S' = \frac{S - \mu_{imp}}{\sigma_{imp}} \tag{3}$$

In the particular context of speaker tracking, the impostor distribution is estimated using all the other known speakers with the same gender as the tracked speaker.

5.6. Decision

The decision strategy relies on an open set speaker identification decision. For a given segment, only the speaker with the best score is retained (others are rejected); i.e. only one decision is made (accepted/rejected) for a speaker and for a segment. A gender independent threshold is applied in order to accept or reject the speaker. The threshold was estimated on the development corpus and is gender independent.

5.7. Proposed speaker detection system

The proposed speaker detection system consists in a fusion between two sub-systems using an arithmetic mean between scores.

The two sub-systems use the same structure, described in previous paragraphs; the difference consists in the data used to train the world model. The sub-system S1 uses gender dependent world models trained on studio condition (S) whereas S2 uses gender dependent world models trained on telephonic condition (TS).

6. Speaker tracking evaluation during ESTER 2005 campaign

For this task, a list of 279 speakers with at least 2 minutes of speech in the training set was provided. The task consisted in tracking each speaker among 10h of test signals.

As said in introduction, two different strategies to link the speaker diarization process and the speaker detection one were evaluated in the framework of the ESTER 2005 campaign. These two strategies rely on:

- the only use of the speaker boundaries provided by speaker diarization (A-SES). The speaker detection decision is made on each segment independently of the others.
- the use of all the speaker information provided by speaker diarization (boundaries and clustering) (A-SRL). The speaker detection decision is made by averaging the scores (per tracked speaker) of all the segments attributed to the same speaker by the speaker diarization system.

Table 3 provides the results in terms of F-Measure for the A-SES strategy. The results of the two sub-systems S1 and S2, are shown as well as the fusion, which improves the performance.

Table 3: Speaker tracking systems results in terms of F-Measure for two sub-systems (S1 and S2) and for their fusion (Main system) on the development (*Dev*) and test (*Eva*) corpus

Corpus	Speaker tracking system		
	S1	S2	Main system (fusion)
Dev	0.629	0.597	**0.633**
Eva	0.605	0.585	**0.659**

For evaluating the influence of automatic speaker turn detection error (i.e. boundary detection error) on speaker tracking performance, we compare, in Table 4, the results of the main system A-SES strategy with the same system using the reference boundaries (Ref-SES). A similar comparison is provided in the same table for the A-SRL strategy. In this case, all the speaker information (boundaries and clustering) automatically provided by the speaker diarization system is compared with the reference speaker information.

As expected, the global performance decreases when using the automatic segmentation compared with the reference one for both the strategies. The observed loss is mainly due to the automatic speaker diarization errors. Indeed, for the strategy A-SES, error analysis has shown that most of these errors are due to gender misclassification (acoustic macro-class segmentation) while for the strategy A-SRL, due to gender misclassification and speaker clustering error. This error cumulation for the latter may explain the difference in performance between the both strategies despite the fact that A-SRL strategy should bring more relevant information to the speaker tracking system. This hypothesis besides is confirmed by the performance of Ref-SRL system which is better than the Ref-SES system.

7. Conclusions

In this paper, we proposed a complete speaker tracking system for broadcast news data. The system is composed of three steps: an acoustic macro-class segmentation, followed by a speaker diarization and a speaker detection processes.

Two strategies to link speaker diarization system to speaker detection has been proposed and evaluated. The acoustic segmentation and the speaker diarization are firstly evaluated separately in order to show their intrinsic performance. The first strategy A-SES uses only speaker boundaries like input for speaker detection while A-SRL strategy uses both speaker clustering information and speaker boundaries. We have analyzed the influence of the two first step errors on the speaker detection process by comparing their performance with this obtained by speaker detection while reference segmentation is used. We showed that using a speaker clustering process can improve the results, even if a small loss was noticed when automatic diarization is used.

Future work will focus on the speaker detection process, which was not tuned for speaker tracking during this work. All the software used in this paper are available thanks to an open software licence.

Table 4: Speaker tracking system results for the two link strategies with automatic segmentation information and with references

Input	F-Measure
A-SES	0.659
Ref-SES	0.752
A-SRL	0.614
Ref-SRL	0.774

8. Acknowledgements

The ALIZE library has been developed by Frederic Wills at LIA laboratory in the framework of the project Technolangue/AGILE/ALIZE, financed by French research ministry.

9. References

[1] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, G. Gravier, "The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News", submited to Interspeech 2005, Lisbon, Portugal,2005.

[2] http://www.afcp-parole.org/workshop.htm.

[3] J.-F. Bonastre, F. Wills, S. Meignier, "ALIZE, a free Toolkit for Speaker Recognition", Proc. of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia,USA, 2005.

[4] http://www.lia.univ-avignon.fr/heberges/ALIZE.

[5] "Spring 2003 Rich Transcription Workshop", 2003.

[6] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J.-F. Bonastre, "The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation", Proc. of International Conference on Acoustics Speech and Signal Processing (ICASSP 2004), Montreal, Canada, 2004.

[7] S. Meignier, J.-F. Bonastre, C. Fredouille, T. Merlin, "Evolutive HMM for speaker tracking system", Proc. of International Conference on Acoustics Speech and Signal Processing (ICASSP 2000), Istanbul, Turkey, 2000, pp. 1177-1180.

[8] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, Y. Magrin-Chagnolleau, "The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation", Proc. of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003), Hong Kong, 2003, Vol. II, pp. 89-92.

[9] J.L. Gauvain, C.H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains", IEEE Trans. Speech and Audio Proc., 22:291–298, 1994.

[10] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, “Score normalization for text-independent speaker verification system”, DSPNIST99, 10(1-3):42–54, 2000.