SPEAKER TRACKING BY ANCHOR MODELS USING SPEAKER SEGMENT CLUSTER INFORMATION

Mikaël Collet [1][2], Delphine Charlet [1], Frédéric Bimbot [2]

(1) France Telecom R&D - TECH/SSTP - 2 av. Pierre Marzin 22307 Lannion Cedex - FRANCE
{mikael.collet, delphine.charlet}@francetelecom.com
(2) IRISA (CNRS & INRIA) - Campus de Beaulieu - 35042 Rennes Cedex - FRANCE
bimbot@irisa.fr

ABSTRACT

In this paper, we present a speaker tracking system entirely based on anchor models approach. The aim of this article is to evaluate if the probabilistic anchor models approach, which models a speaker by a normal distribution in the anchor models space, gives good performances in speaker tracking and also to investigate how speaker segment cluster information can improve speaker tracking performances. Evaluation is done on the audio database of the ESTER evaluation campaign for the rich transcription of French broadcast news. Results show that deterministic metrics on anchor models are suitable for segmentation and clustering tasks, whereas the probabilistic approach on anchor models gives interesting results for speaker-tracking. It is also observed that tracking performances are improved when all segments of a cluster are pooled together prior to the classification process. This improvement manifests itself as an improved recall rate on short segments.

1. INTRODUCTION

The anchor models approach for speaker recognition consists in modelling a speaker relatively to a fixed set of known speakers. It is particularly suitable to the task of speaker indexing in large audio databases [1], because the main computational burden can be performed off-line, independently of the target speaker. Then, for any new target speaker to be tracked, the additional computation is very low. Anchors models can be seen as a speech signal representation (such as cepstral features), but particularly suitable to speaker modelling. In a previous paper [2], a speaker tracking system using anchor models was proposed and achieved interesting performances, when including the target speaker into the anchor models space and performing selection of anchor models according to the target speaker. In this paper, we want to investigate speaker tracking with an anchor models space totally independent from the target speaker, strictly in the sense of speech signal representation, thus excluding the target speaker and without any anchor speaker selection.

The proposed speaker tracking system is a 3 step process: segmentation, clustering and detection. All these steps are based on Anchor Models.

The paper is organized as follows: in section 2, we briefly recall the concept of Anchor Models. Then, in section 3, the use of a deterministic metric between representations in the anchor model space is proposed to perform speaker segmentation and clustering. In section 4, a probabilistic approach on Anchor Models is used to perform speaker detection, and a detection measure that uses speaker clustering is proposed. Finally, an evaluation on the French broadcast news database of the ESTER evaluation campaign is presented in section 5.

2. CONCEPT OF ANCHOR MODELS

Recent research [1][2] have been oriented on relative speaker representation. This modelling consists in projecting a speaker utterance into a space of reference speakers. The speaker is not represented in an absolute way but relatively to a set of speakers whose models (for instance GMM-UBMs) are trained beforehand. These models are called anchor models.
The speaker is characterized by the vector of the log likelihood ratio between the speaker data and the anchor models. This vector is called Speaker Characterization Vector (SCV) and is denoted $\widetilde{X}$.

$$\widetilde{X} = \begin{bmatrix} \widehat{s}(X|\overline{\lambda}_1) \\ \widehat{s}(X|\overline{\lambda}_2) \\ \vdots \\ \widehat{s}(X|\overline{\lambda}_E) \end{bmatrix} \quad (1)$$

where $\widehat{s}(X|\overline{\lambda}_e)$ is the average log likelihood ratio of the data X (of N acoustics feature vectors) for the GMM model of the reference speaker $\overline{\lambda}_e$ relatively to a Universal Background Model:

$$\widehat{s}(X|\overline{\lambda}_e) = \frac{1}{N} \log \frac{p(X|\overline{\lambda}_e)}{p(X|\lambda_{UBM})} \quad (2)$$

where λ_{UBM} is the Universal Background Model which is used to initialize the training of the anchor models associated to reference speakers.
In this paper, we propose to investigate on the behaviour of this modelling for the tasks of speaker segmentation, speaker clustering and speaker detection.

3. SPEAKER SEGMENTATION AND CLUSTERING USING ANCHOR MODELS

In this section, we focus on how to use the anchor models approach for speaker segmentation and clustering tasks as an alternative to the classical mono-gaussian modelling in the acoustic space [3].

3.1. Speaker segmentation

The speaker segmentation process consists in segmenting an audio document in homogeneous segments of reasonable length which are assumed to have been pronounced by only one speaker. This task of speaker segmentation is carried out with no prior knowledge on the speaker(s) to be detected in a later stage of speaker tracking.
A commonly used technique consists in detecting some statistical ruptures in the signal corresponding to a speaker change [3]. This method computes a score criterion along the speech signal and then detects ruptures.
The first step of the speaker segmentation process consists in calculating a measure of similarity between two successive segments $A = \{y_{t-T}...y_{t-1}\}$ and $B = \{y_t...y_{t+T-1}\}$ where each segment has in practice a length of 2.4 s. The window composed of the two segments is shifted every 160 ms along the speech signal and at each shift a measure is calculated. Classically, this measure is computed as the Generalized Likelihood Ratio (GLR) based on mono-gaussian modelling of speech segments [3]. In our system, the two consecutive segments are represented by their SCV in the anchor models space ($\widetilde{A}$ and $\widetilde{B}$) and a correlation metric (refered to in previous works as the deterministic approach [4]) is computed between the two SCV:

$$\rho(\widetilde{A}, \widetilde{B}) = 1 - R(a, b) \tag{3}$$

$$R(a, b) = \frac{C_{ab}}{\sigma_a \sigma_b} \tag{4}$$

The components of the two SCV $\widetilde{A}$ and $\widetilde{B}$ can be considered as realisations of two random variables a and b, C_{ab} is the covariance between the two variables and σ_a, σ_b are respectively the standard deviation of a and b.
This process gives as output a score criterion where the most significant local maxima are considered as a speaker change.
The local maxima detection is performed according to the method proposed in [5], where a decision threshold is set so as to get the same mean segment length in each audio document.

3.2. Speaker clustering

Speaker clustering consists in grouping into a same cluster segments which are supposed to be pronounced by a same speaker with no prior information on the speakers of the audio document. A large variety of clustering algorithms have been investigated in different contexts.
The clustering algorithm used in this work is based on a single-linkage approach and can be performed in three steps :

1. To compute a measure of similarity between all the segments as in equation 3.
2. To group into a same cluster a segment and its nearest neighbour according to the measure of similarity.
3. To merge all the clusters of two segments whose intersection is not empty.

Compared to a bottom-up clustering algorithm, the single linkage technique presents the advantage that no threshold needs to be tuned to stop the clustering process.
As for the speaker segmentation approach presented in the previous section, the speaker clustering step is based on the correlation measure of equation (3).

4. SPEAKER DETECTION USING THE PROBABILISTIC ANCHOR MODELS APPROACH

In this section, we present a speaker detection system using the probabilistic anchor models approach proposed in [4]. This approach models a speaker by a normal distribution of that speaker's SCV in the anchor models space. Doing this, a priori information can be used for speaker modelling and a probabilistic metric, that models intra-speaker variability can be applied between the test occurrence and speakers models.
We also investigate how speaker segment cluster information can improve speaker detection performances [6] and propose a new measure of similarity between a target speaker and a test segment according to the cluster which the segment belongs to.

4.1. Speaker detection process

The speaker detection process using the probabilistic anchor models approach computes a likelihood score between the test segment SCV $\widetilde{X}$ and the target speaker model S in the anchor space. The score is computed as a likelihood ratio between the target speaker's SCV model and a speaker independant gaussian model of the SCV:

$$L(\widetilde{X}|S) = \log \frac{p(\widetilde{X}|\mu_S, \Sigma_0)}{p(\widetilde{X}|\mu_0, \Sigma_0)} \tag{5}$$

The covariance matrix is the same for all speaker models and is equal to Σ_0. The mean vector of a speaker model is adapted from the mean vector μ_0 with a simplified version of MAP.
The a priori distribution parameters (Σ_0 and μ_0) are estimated from a development set according to the process described in [4].
The probabilistic approach also considers the segment X as a model and the target speaker S as a test occurrence and a symmetric probabilistic metric is defined as :

$$L(\widetilde{X}, \widetilde{S}) = \frac{L(\widetilde{X}|\widetilde{S}) + L(\widetilde{S}|\widetilde{X})}{2} \tag{6}$$

4.2. Speaker detection using speaker segment cluster information

The aim of the approach presented in this section is to build clusters of speaker segments. By grouping together similar segments before speaker detection, we expect to enhance the robustness of the classification process. In particular, we expect that the measure of similarity computed for classifying short segments will be more reliable and therefore short segments will be better detected by the speaker detection process. The limitations of such an approach is that speaker tracking performances can be deteriorated by clustering errors which decrease the average cluster purity (ACP).
This section proposes a solution for using the clustering information in the speaker tracking system. The measure of similarity between a segment X and a target speaker S is replaced by a measure of similarity computed between the target speaker S and every segment of the cluster C_X which the segment X belongs to.
The measure of similarity between X and S according to the cluster C_X is defined as :

$$L(\widetilde{X}, \widetilde{S}) \rightarrow L_{C_X}(\widetilde{X}, \widetilde{S}) = \frac{1}{N_{C_X}} \sum_{i=1}^{N_{C_X}} L(\widetilde{Y}_i, \widetilde{S}) \tag{7}$$

where N_{C_X} is the number of segments Y_i of the cluster C_X which the segment X belongs to.

5. EXPERIMENTS AND RESULTS

The speaker tracking system based on anchor models is evaluated on the speaker tracking task of the French ESTER broadcast news evaluation campaign [7]. The evaluation corpus, the evaluation measure and the system configuration are presented in the following sections before giving results.

5.1. Evaluation corpus

The corpus used for this experiment is a corpus of radio broadcast news in french. The corpus is divided into a training set, a development set and a test set, according to the ESTER phase 2 specifications (see [7] for details). The training set contains 82h of broadcast news, recorded during the period of 1998-2003. The development set contains 10h of broadcast news, recorded in 2003 and the test set 10h of broadcast news, recorded in 2004, corresponding to the same radio stations as the development set plus two additional radio stations. According to [6], the development set is divided in two sets (dev1 and dev2) and the experiments presented in this paper are done on the set dev1 with a list of 279 target speakers provided by the ESTER organizer.

5.2. Evaluation measure

5.2.1. Speaker segmentation and clustering evaluation measure

In our experiments, speaker segmentation and clustering performances are evaluated in term of Average Cluster Purity (ACP) [8] according to the mean cluster length. The Average Cluster Purity is defined as :

$$ACP = \frac{1}{N_0} \sum_{i=1}^{C} \max_{j=1:S} n_{ij} \tag{8}$$

where N_0 is the number of frames from the audio document, C the number of clusters, S the number of speakers and n_{ij} the number of frames from the cluster i pronounced by the speaker j.

5.2.2. Speaker tracking evaluation measures

Speaker tracking performance is evaluated in terms of Precision/Recall where Precision (PR) and Recall (RC) are defined by :

- $PR = \frac{\text{Target speaker duration detected}}{\text{Duration detected}}$
- $RC = \frac{\text{Target speaker duration detected}}{\text{Target speaker duration}}$

The Precision and Recall values are combined in a single evaluation measure using the common F-measure:

$$F = \frac{2.PR.RC}{PR + RC} \tag{9}$$

5.3. System configuration

In all experiments, 13 Mel-frequency cepstral coefficients with their first and second derivatives plus ΔE and $\Delta\Delta E$ are used and the anchor models are 256-component GMMs adapted from a UBM model with a MAP criterion. The anchor models space is compounded of all the speakers, different from target speakers, which have more than 70 seconds of speech available in the training set: they are 316. Feature warping [9] is applied for the speaker detection task, whereas no channel compensation is performed for speaker segmentation and clustering.

5.4. Results

5.4.1. Speaker segmentation and clustering performances

Table 1 presents speaker segmentation performances for 2 types of distances between consecutive segments of speech: GLR and correlation on anchor models (AM). Results on speaker clustering using anchor models are also presented. Performances are evaluated in terms of ACP for the corresponding mean segment length or mean cluster length.
Results show that for a same mean segment length, the anchor models approach gives a better ACP in speaker segmentation than the GLR. Although speaker clustering errors impair the ACP (90.1 % against 96.7 %) however, the mean cluster length is significantly increased (68.7 s against 12 s).

	ACP	Mean length
Speaker segmentation GLR	93.4	12.0
Speaker segmentation AM	96.7	12.0
Speaker clustering AM	90.1	68.7

Table 1. Speaker segmentation and clustering performances

5.4.2. Speaker tracking performances

Four systems of speaker tracking are compared and their performance is depicted on Figure 1 (in terms of Precision/Recall) and on Figure 2 (in terms of DET Curves) :
- Tracking using the deterministic approach (as in [4]).
- Tracking using the deterministic approach with clustering.
- Tracking using the probabilistic approach.
- Tracking using the probabilistic approach with clustering.

In systems using the deterministic approach, the measure of similarity between a segment and a target speaker defined by equation 5 is replaced by the correlation metric defined by equation 3.
As can be seen on these figures, the probabilistic anchor models approach gives good results in speaker tracking and yields a significant improvement over the deterministic anchor models approach. These results confirm those reported in [4] for a speaker verification task and are similar to those reported in [6] [10] with direct GMM modelling. These figures also show that speaker tracking performances are better when speaker clustering is performed.
In order to explain these results, we analyse tracking performances at the optimal operating point (point of the curve which maximizes the F-Measure) for each system. Optimal operating points are reported in Table 2. This table indicates that for a same precision rate (51.7 and 50.7 for the deterministic approach, 89.2 and 89.6 for the probabilistic approach), the recall rate for tracking with clustering is significantly better than the recall rate for tracking without clustering. In terms of Missed Detections, the clustering gives a relative error rate reduction of 14 % for the deterministic approach and 35 % for the probabilistic approach.
These results indicate that more segments are correctly detected when clustering is performed. An analysis shows that 80% of these segments are shorter than 8 seconds. This analysis confirms our assumption that short segments are better characterized with the clustering process and therefore they are better detected by the tracking system.

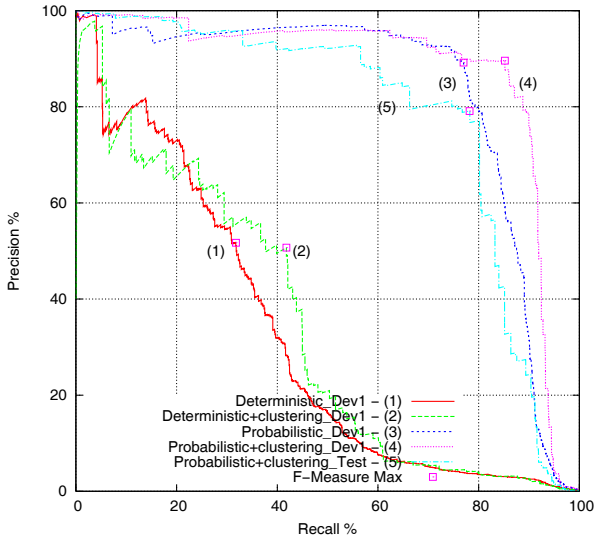

Fig. 1. Precision versus Recall for the 4 speaker tracking systems

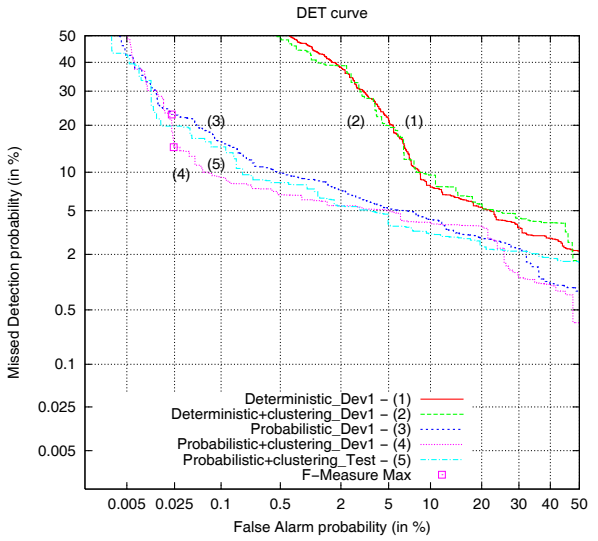

Fig. 2. Missed Detections versus False Alarms for the 4 speaker tracking systems

Set	Approach	F-Max	PR	RC
Dev1	Deterministic	39.4	51.7	31.8
Dev1	Deterministic. + clustering	45.8	50.7	41.8
Dev1	Probabilistic	82.7	89.2	77.0
Dev1	Probabilistic + clustering	87.4	89.6	85.2
Test	Probabilistic + clustering	78.7	79.1	78.2

Table 2. Speaker tracking systems optimal operating points

6. CONCLUSION

In this article, we have presented a speaker tracking system entirely based on a concept of a relative speaker representation. This concept consists in representing a speaker in a space of reference speakers called anchor models. This anchor models concept using the correlation metric applied to speaker segmentation gives better performances than a mono-gaussian modelling using a GLR measure of similarity in the acoustic space.
Then, the anchor models concept using the probabilistic approach is used for speaker tracking and achieves promising results, confirming previous results obtained for speaker verification on telephone speech database [4].
Finally, we have also investigated how to use clustering information in a speaker tracking system. A new measure of similarity derived from the measure of similarity between a target speaker and all the segments of a cluster is proposed. Evaluations show that for a same precision rate, the recall rate is significantly improved by using clustering information. Grouping similar segments into a same cluster before speaker detection tends to yield a more robust and reliable characterization of individual segments, especially short ones.

7. REFERENCES

[1] D.E. Sturim, D.A. Reynolds, E. Singer, and J.P. Campbell, "Speaker indexing in large audio databases using anchor models," in *ICASSP2001*, 2001, pp. 429–432.

[2] M. Collet, D. Charlet, and F. Bimbot, "A correlation metric for speaker tracking using anchor models," in *ICASSP*, 2005.

[3] P. Delacourt, D. Kryze, and C. Wellekens, "Speaker-based segmentation for audio data indexing," in *ESCA ETRW Workshop*, 1999.

[4] M. Collet, Y. Mami, D. Charlet, and F. Bimbot, "Probabilistic anchor models approach for speaker verification," in *INTERSPEECH*, 2005.

[5] M. Seck, R. Blouet, and F. Bimbot, "The IRISA/ELISA speaker detection and tracking systems for the NIST'99 evaluation campaign," *Digital Signal Processing*, vol. 10, pp. 154–171, 2000.

[6] D. Moraru, M. Ben, and G. Gravier, "Experiments on tracking and segmentation in radio broadcast news," in *INTERSPEECH*, 2005.

[7] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, G. Gravier, and J-F. Bonastre, "The ESTER phase 2 evaluation campaign for the rich transcription of french broadcast news," in *INTERSPEECH*, 2005.

[8] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1998, pp. 757–760.

[9] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *proc of A Speaker Odyssey*, 2001, number 1038.

[10] D. Isastre, N. Scheffer, C. Fredouille, and J.F. Bonastre, "Broadcast news speaker tracking for ESTER 2005 campaign," in *INTERSPEECH*, 2005.