A CORRELATION METRIC FOR SPEAKER TRACKING USING ANCHOR MODELS

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}@rd.francetelecom.com
(2) IRISA (CNRS & INRIA) - Campus de Beaulieu - 35042 Rennes Cedex - FRANCE
bimbot@irisa.fr

ABSTRACT

This paper presents an approach for speaker tracking in a large audio database. The system described is based on a speaker segmentation procedure consisting in a detection of statistical ruptures in the speech signal followed by a speaker detection procedure using anchor models. The technique of anchor modelling is presented and a new metric to compare speech segment based on the correlation coefficient is introduced. This novel metric is evaluated and compared to the classical Euclidean and Angular metrics for the speaker detection task. Evaluation are done on the audio database of the ESTER evaluation campaign for the rich transcription of French broadcast news. The new metric appears to be more efficient than the classical metrics for the task of speaker detection.

1. INTRODUCTION

Since recent years, a lot of audio data (like broadcasts news) are stored in large databases. In this context, the task of speaker tracking, consisting in searching speech utterances of a target speaker, becomes difficult. Actually, the important size of audio archives increases the computing time of the speaker tracking system and therefore it limits its performances in a real time application. In the literature, two main approaches for speaker tracking are proposed. The first one consists in segmenting the audio signal and then detecting the target speaker [1]. In the second approach, the segmentation and the detection are done simultaneously [2].
The speaker tracking system proposed in this paper, is based on the first approach and is composed of two modules. The first one, detailed in section 2, consists in segmenting the audio signal in portions which are assumed to have been pronounced by only one speaker. This task of speaker segmentation is done off-line. The second one consists in detecting speech utterances of the target speaker. This speaker detection module described in section 3 compares the target speaker uterrances with all the segments from the speaker segmentation module and decides if segments have been pronounced by the target speaker. This process uses the anchor models technique [3] to modelize all the speech utterances (target speaker and speech segments) by a characteristic vector independently of the utterance length. This technique reduces the size of speech segments models and therefore the size needed for representing the audio databases. In the last section of the paper, the speaker tracking system is evaluated on the ESTER 2003 evaluation corpus and the classical metric used in the anchor modelling technique are compared to the new metric proposed in section 3.3.

2. SPEAKER SEGMENTATION

The first step of the system consists in segmenting the audio document in homogeneous segments of reasonable length and 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 the next step. The most used technique consists in detecting some statistical ruptures in the signal corresponding to a speaker change. This method computes a score criterion along the speech signal and then detects ruptures.

2.1. Score criterion computation

In a context of audio data indexing, [4] describes a speaker segmentation system based on a score criterion computation. This classical technique, used for speaker tracking, consists in calculating a statistical distance between two consecutive segments $a = \{y_1...y_{t-1}\}$ and $b = \{y_t...y_T\}$ where each segment is of length 2.4 s. The window composed of the two segments is shifted every 160 ms along the speech signal and at each shift a distance derived from the Generalized Likelihood Ratio (GLR)[5] is calculated.

$$R = \frac{L(ab, \hat{\mu}_{ab}; \hat{\Sigma}_{ab})}{L(a, \hat{\mu}_a; \hat{\Sigma}_a) L(b, \hat{\mu}_b; \hat{\Sigma}_b)} \tag{1}$$

where $L(x, \hat{\mu}_x; \hat{\Sigma}_x)$ represents the likelihood of the acoustic sequence x for the multi-gaussian process $\mathcal{N}(\mu_x; \Sigma_x)$ and ab is the concatenation of utterances a and b.
The GLR distance is computed by taking the logarithm of the previous expression :

$$d_{GLR} = -log(R) \tag{2}$$

This process gives as output a score criterion where the most significant local maxima are considered as statistical breakpoints.

2.2. Statistical breakpoint detection

The second step of the speaker segmentation process is to detect the local maxima of the score criterion. The statistical breakpoint detection method proposed in [6] is used in the system. This method computes a breakpoint criterion which can be defined as the magnitude of a local maxima relative to the highest of the two surrounding minima on each side of that extrema and dominated by it.
If the breakpoint criterion is higher than a threshold then a statistical breakpoint is detected. This threshold permits to tune the

0-7803-8874-7/05/$20.00 ©2005 IEEE

number of segments at the output of the speaker segmentation system but the optimal threshold can be different from one document to another.
To cope with this problem, the threshold is determined a posteriori in order to have the same segment mean length for each document.

3. SPEAKER DETECTION USING ANCHOR MODELS

Speaker detection systems proposed in the literature are based on a GMM-UBM modelling of speakers. In this article, a system using the anchor modelling technique is proposed.

3.1. Concept of anchor models

Recent research [3][7] have been oriented on a new 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 GMM-UBM models are pre-trained. These models are called anchor models.
The speaker is characterized by a vector defined as the set of the likelihood between the speaker data and the anchor models. This vector is called Speaker Characterization Vector (SCV) and denoted $\widetilde{X}$.

$$\widetilde{X} = \begin{bmatrix} \widehat{s}(X|\overline{\lambda}_1) \\ \widehat{s}(X|\overline{\lambda}_2) \\ \vdots \\ \widehat{s}(X|\overline{\lambda}_E) \end{bmatrix} \tag{3}$$

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 $\overline{\lambda}_e$ relative to a Universal Background Model :

$$\widehat{s}(X|\overline{\lambda}_e) = \frac{1}{N} \log \frac{p(X|\overline{\lambda}_e)}{p(X|\lambda_{UBM})} \tag{4}$$

where λ_{UBM} being the Universal Background Model which has been used to initialize the training of the anchor models.
With this modelling, the speaker detection step can be viewed as projecting the target speaker and all the unknown segments into the anchor space. Then a metric between the speaker and each segment is calculated and finally, the metric is compared to a threshold to decide whether the segment has been uttered by the target speaker.

3.2. Anchor models selection

This representation needs a set of GMM models to create a reference space for the speaker utterance projection.
For a task of speaker indexing in an audio document, [8] proposes to choose E speakers among a set of unknown speakers which maximize the likelihood of the whole data of the document. This method matches the anchor models to the speech document.
For the task of speaker detection, it is assumed that the anchor models have to be matched to the target speaker. So for each speaker to detect, a Speaker Characterization Space (SCS) is created by choosing the $E = 50$ nearest speakers to the target among a set of 597 speakers. In this system, the target speaker model is always included in the SCS. Therefore the 49 other models can be considered as support models to the target speaker.
The number of models has been optimized by a preliminary experiment which shows that $E = 50$ gives the best performance.

3.3. Metric for SCV comparison

The classical metrics for SCV comparison are the Euclidean metric and the Angular metric [7]. The efficiency of a metric depends on its capacity to be robust against the mismatch (recording condition, intra-speaker variability) between the training data and the testing data. This section first details the classical metrics and the kind of mismatch they are robust against. Then a new metric based on the correlation coefficient is described.

Let X and Y two speech segments, $\widetilde{X}$ and $\widetilde{Y}$ their Speaker Characterization Vector.

- Euclidean metric :

$$d(\widetilde{X},\widetilde{Y}) = \sqrt{|\widetilde{X}-\widetilde{Y}|^2} \tag{5}$$

with

$$d(\widetilde{X},\widetilde{Y}) = 0 \Longleftrightarrow \widetilde{Y} = \widetilde{X} \tag{6}$$

This metric is efficient when there is no mismatch between the training data and the testing data.

- Angular metric :

$$\delta(\widetilde{X},\widetilde{Y}) = \arccos \left[\frac{\widetilde{X}\widetilde{Y}^T}{\sqrt{\widetilde{X}\widetilde{X}^T.\widetilde{Y}\widetilde{Y}^T}} \right] \tag{7}$$

with

$$\delta(\widetilde{X},\widetilde{Y}) = 0 \Longleftrightarrow \widetilde{Y} = a\widetilde{X} \quad \forall a \in \Re \tag{8}$$

Thanks to this property, the angular metric is robust against a mismatch modelized by a multiplicative coefficient a between the two SCV.

- New metric : Figure 1 shows the relation between the components of two SCV (from a same speaker on the left and from differents speakers on the right). Thanks to this experimental observation, it is assumed that an affine relation exists between two SCV from the same speaker.

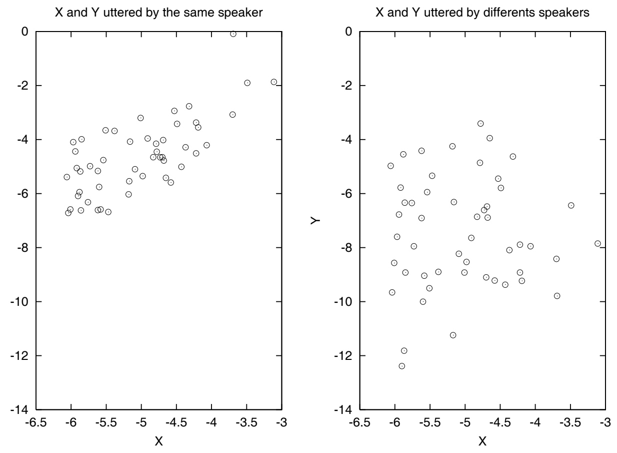

Fig. 1. Relation between SCV components

Therefore, it is desirable that the new metric satisfies the property :

$$\rho(\widetilde{X},\widetilde{Y}) = 0 \Longleftrightarrow \widetilde{Y} = a\widetilde{X} + b \quad \forall (a,b) \in \Re^2 \tag{9}$$

So, if the components of the two SCV are considered as the realisation of two random variables x and y, this property can be evaluated by the correlation coefficient $R(x, y)$ [9] :

$$R(x, y) = \frac{C_{xy}}{\sigma_x \sigma_y} \tag{10}$$

where C_{xy} is the covariance between the two variables and σ_x, σ_y are respectively the standard deviation of x and y. Correlation coefficient properties and experimental observations permit to define a new metric to compare SCV for speaker detection :

$$\rho(\widetilde{X}, \widetilde{Y}) = 1 - R(x, y) \tag{11}$$

Therefore, the relation $\rho(a\widetilde{X} + b, \widetilde{Y}) = \rho(\widetilde{X}, \widetilde{Y})$ is always true for $a > 0$ and this metric is robust against a mismatch that is modelized by a positive multiplicative coefficient and an additive coefficient.

At first sight, these three metrics appear to be very different, but a relation can be established between the Euclidean metric and the two other metrics. This relation depends on the kind of normalization that is applied to SCV. The two kinds of normalization are the euclidean normalization and the centering-reduction normalization which have been already used by [8].

- Euclidean normalization : $\widetilde{X}_N = \frac{\widetilde{X}}{\sqrt{\widetilde{X}\widetilde{X}^T}}$
  After normalization, the Euclidean metric can be expressed by a monotonic function of the Angular metric :

$$d_N(\widetilde{X}, \widetilde{Y}) = \sqrt{2}\sqrt{1 - \cos\left[\delta(\widetilde{X}, \widetilde{Y})\right]} \tag{12}$$

- Centering-reduction normalization : $\widetilde{X}_{CR} = \frac{\widetilde{X} - \mu_{\widetilde{X}}}{\sigma_{\widetilde{X}}}$
  After normalization, the Euclidean metric can be expressed by a monotonic function of the correlation metric :

$$d_{CR}(\widetilde{X}, \widetilde{Y}) = \sqrt{2E}\sqrt{\rho(\widetilde{X}, \widetilde{Y})} \tag{13}$$

4. EXPERIMENTS AND RESULTS

The speaker tracking system based on anchor models described in this paper was evaluated on the speaker tracking task of the French ESTER broadcast news evaluation campaign [10]. The three metrics previously detailed are compared to a GMM-UBM speaker detection system as in [3].
The evaluation corpus, the evaluation measure and the system configuration are presented in the following sections before giving results.

4.1. 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 1 specifications (see [10] for details). The training set contains 38 broadcasts corresponding to 19h40 of France-Inter (Inter, 27 broadcasts) and 11h of Radio France International (RFI, 11 broadcasts). The developement set and the test set contain six broadcasts corresponding to 2h40 of France-Inter (Inter, 4 broadcasts) and 2h of Radio France International (RFI, 2 broadcasts). The speaker tracking is performed on the development set independently for each file and each target speaker.
For the experiments on the development set, 91 target speakers are used and 96 on the test set.

4.2. Evaluation measure

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

- $PR = \frac{\text{Target speaker time detected}}{\text{Time detected}}$
- $RC = \frac{\text{Target speaker time detected}}{\text{Target speaker time}}$

The Precision and Recall values are combined in a single evaluation measure using the common $F - measure$ [11], which is defined as :

$$F = \frac{2.PR.RC}{PR + RC} \tag{14}$$

4.3. 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 statistical models are 256-component GMMs. CMS was applied. The 597 models used for the speaker space selection are adapted from a UBM model with a MAP criterion with data from the training set. The speaker models used in the GMM-UBM system are also adapted with a MAP criterion but from a gender-dependent UBM.

4.4. Results

4.4.1. *Influence of the metric*

Figure 2 shows the Precision/Recall trade-off for the GMM-UBM system, correlation anchor model, euclidean anchor model and angular anchor model systems obtained with the manual segmentation (the manual segmentation correspond to breath group segmentation with a segment mean length of about 4 s). Each points of these curves corresponds to a different detection threshold. The operating points of each systems which maximize the $F-measure$ are summarized in Table 1 and are also marked by a '+' on figure 2.
According to the $F-measure$, the correlation anchor model system gives the best performance and yields a significant improvement over the other systems. This result confirmes the assumption that mismacth is better modelized by a multiplicative coefficient *and* an additive coefficient. Experiments without CMS were also conducted and give the same conclusion.

System	F_{max}	RC_{max}	PR_{max}
GMM_UBM	65.1	62.5	67.8
Angular Anchor Model	60.4	50.8	74.3
Euclidean Anchor Model	66.5	59.2	75.9
Correlation Anchor Model	78.6	72.7	85.6

Table 1. Operating points for the manual segmentation

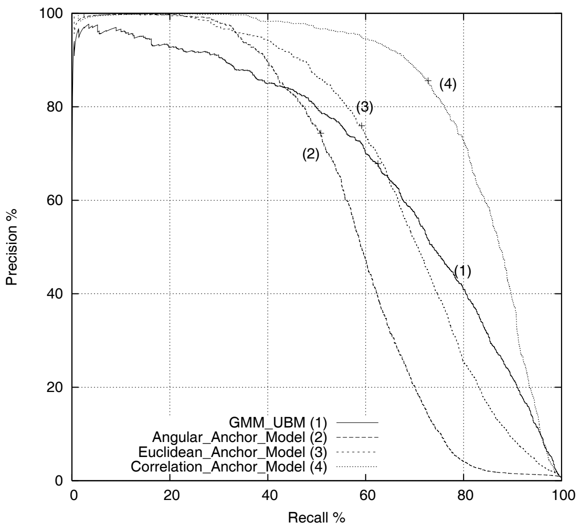

Fig. 2. Precision versus Recall for the GMM-UBM and anchor model systems for the manual segmentation

4.4.2. *Influence of the segmentation*

Figure 3 represents the influence of the segmentation on the performance of the correlation anchor model system and shows that the system gives the best performance for a segment mean length of 10 s.
The difference between the manual segmentation and the best segmentation is due to the fact that our system shows a better recall performance on long segments than on short segments.
The effect of the segmentation error (segments are not homogeneous) can be seen when the manual segmentation and the 4 s segments mean length segmentation are compared. In this case, the segment mean length are the same but the non-homogeneity of the segments deteriorates the system performance.

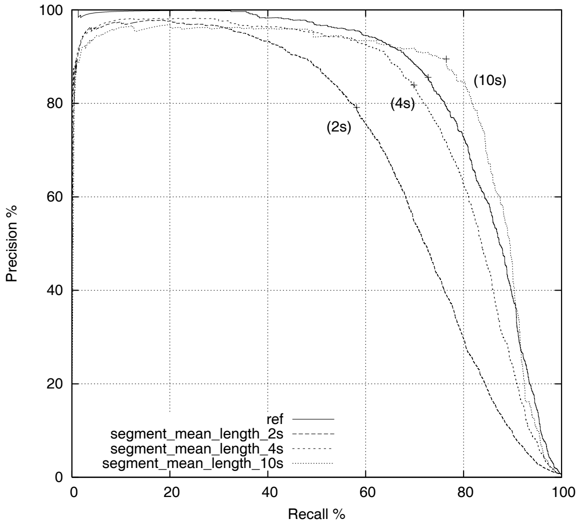

Fig. 3. Precision versus Recall for several segmentation for the correlation anchor model system

5. CONCLUSION

This paper presented a speaker tracking system based on anchor models. In a first part, a classical speaker segmentation system was described. Then, the speaker modelisation using anchor models was introduced and a new metric derived from the correlation coefficient has been proposed. This metric appears to be more robust against mismatch between training data and testing data than the classical euclidean and angular metric. Therefore, in the future, research efforts will be focused on using this new metric for the task of speaker segmentation in order to include the speaker segmentation module in the speaker detection module.

6. REFERENCES

[1] Lie Lu and Hong-Jiang Zhang, "Speaker change detection and tracking in real-time news broadcasting analysis," in *ACM International Conference on Multimedia*, 2002, pp. 602–610.

[2] I-M. Chagnolleau, A-E. Rosenberg, and S. Parthasarathy, "Detection of target speakers in audio databases," in *ICASSP'99*, 1999.

[3] 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.

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

[5] H. Gish, M-H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1991, pp. 873–876.

[6] 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.

[7] Yassine Mami and Delphine Charlet, "Speaker identification by location in an optimal space of anchor models," in *International Conference on Spoken Language Processing*, 2002, vol. 2, p. 1333.

[8] Yuya Akita and Tatsuya Kawahara, "Unsupervised speaker indexing using anchor models and automatic trancription of discussions," in *EUROSPEECH 2003*, 2003, pp. 2985–2988.

[9] Carol Ash, "Correlation," in *The probability tutoring book*, 1993, pp. 235–241.

[10] G. Gravier, J-F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri, "The ester evaluation campaign of rich transcription of french broadcast news," in *Language Evaluation and Resources Conference*, 2004, www.afcp-parole.org/ester.

[11] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for automatic segmentation of audio data," in *ICASSP 2000*, 2000.