"ISI" A New Method for Automatic Speaker Tracking and Detection.

S. Ouamour*, M. Guerti, H. Sayoud**

*USTHB, Electronics Institute, BP 32 Bab-Ezzouar, Alger, Algeria
Email: *ouamour@hotpop.com **sayoud@hotpop.com

Abstract — **In this paper we propose a new algorithm called ISI or "Interlaced Speech Indexing", developed and implemented for the task of speaker detection and tracking. It consists in finding the identity of a well-defined speaker and the moments of his interventions inside an audio document, in order to access rapidly, directly and easily to his speech.**
Speaker Tracking can broadly be divided into two problems: Locating the points of speaker change (Segmentation of the document) and looking for the target speaker in each segment using a verification system in order to extract his global speech in the document: Speaker Detection.
For the segmentation task, we developed a method based on an interlaced equidistant segmentation (IES) associated with the ISI algorithm. This approach uses a speaker identification method based on Second Order Statistical Measures (SOSM). As SOSM measures, we choose the "µGc" one, which is based on the covariance matrix. However, the experiments showed that this method needs, at least, a speech length of 2 seconds, which means that the segmentation resolution will be 2 seconds. By combining the SOSM with the new Indexing technique (ISI), we demonstrate that the average segmentation error is reduced to only 0.5 second, which is more accurate and more interesting for real-time applications.
Results indicate that the association SOSM-ISI provides a high resolution and a high tracking performance: the tracking score (percentage of correctly labelled segments) is 95% on TIMIT database and 92.4% on Hub4 database.

Index Terms — **Speech processing, Speaker tracking, Segmentation.**

I. INTRODUCTION

Speaker tracking and detection consists in finding, in an audio document, all the occurrences of a particular speaker (target). But with the evolution of the information technology and the communications (broadcasting satellite, internet, etc), it exists thousands of television and radio channels which transmit a huge quantity of information. Among this incredible number of information, finding the utterances and their corresponding moments of one particular speaker in an audio document requires that these documents must be properly archived and accessed, for this purpose many existing techniques are using different keys (key word, key topic, etc), however these techniques can be not efficient enough for the task of speaker tracking in audio documents. A more suitable key for these documents could be the speaker identity.
In that sense, the speaker is known a priori by the system (i.e. a model of his features is available in the reference book of the system). Then, the task of tracking can be seen, herein, as a speaker verification task applied locally along a document containing multiple (and unknown) interventions of various speakers: *Speaker Detection*. The Begin/End points of the tracked speaker interventions have to be found during the process. At the end of this process, the different utterances of the tracked speaker are gathered to obtain the global speech of this particular speaker in the whole audio document.
Thus, the research work presented in this paper is set in this context. So, we have developed for this task, a new system based on SOSM measures and a new interlaced speech indexing algorithm (ISI). This algorithm is easy to implement, simple to use and efficient since it has significantly improved the results: In fact, this association has enhanced the scores of speaker tracking with a good segmentation accuracy and good resolution.
So, this paper is organised as follows:
In section 2, we present a brief overview of tracking techniques. Section 3 presents the SOSM-based method. In section 4, we explain the different algorithms introduced for the speaker tracking. Finally, in section 5, we discuss our results and give a global conclusion for this work in section 6.

II. A BRIEF OVERVIEW OF TRACKING TECHNIQUES

In this section, we provide a brief review of existing audiovisual tracking techniques.

In the same field of speaker tracking, Meignier [1] uses a progressive Markov Model (during the indexing process, the model changes at each new speaker detection) and tests it on a subset of the Switchboard database. Magrin-Chagnolleau [2] uses a GMM model for the tracking of one target speaker on Hub4 broadcast news. Cettolo [3] uses a similar approach (i.e a GMM to model each speaker and each generic audio class) on the Italian Broadcast News Corpus. A new tracking scheme presented by Johnson [4] is based on both agglomerative and divisive clustering strategies and is applied on the Hub-4 development data.

Generally, those methods use three types of segmentation: using silence/activity detectors, detecting features change of the speech or identifying the segments nature/type.

In our research work, we have approached the problem with a new method:

We first use an interlaced equidistant segmentation, then a speaker detection technique for the labeling and finally, in case of confusion or transition errors, a new rule for correction and clustering is proposed to ensure that task: all this process is called the ISI process.

ISI represents a good compromise as it is easy to implement, simple to reproduce, inexpensive in computation and it provides high tracking performance.

III. THE SOSM-BASED METHOD

Our speaker identification method, based on mono-Gaussian models [5] [6], uses some measures of similarity which are called Second Order Statistical Measures (SOSM). We recall bellow the most important properties of the approach [5] [6].

Let $\{x_t\}_{1\le t\le M}$ (respectively $\{y_t\}_{1\le t\le N}$) be a sequence of M (respectively N) vectors resulting from the *P*-dimensional acoustic analysis of a speech signal uttered by speaker $\boldsymbol{x}$ (respectively $\boldsymbol{y}$). These vectors are summarized by the mean vector $\overline{x}$ (respectively $\overline{y}$) and the covariance matrix X (respectively Y) :

The similarity measure $\mu_{GC}(\boldsymbol{x},\boldsymbol{y})$ between test utterance $\{y_t\}_{1\le t\le M}$ of speaker $\boldsymbol{y}$ and the model of speaker $\boldsymbol{x}$ is defined by:

$$\mu_{GC}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{P}\left[-\log\left(\frac{\det(Y)}{\det(X)}\right) + tr(YX^{-1})\right] - 1 \quad (1)$$

A symmetric measure can be constructed by combining $\mu_{GC}(\boldsymbol{x},\boldsymbol{y})$ with its dual term $\mu_{GC}(\boldsymbol{y},\boldsymbol{x})$, leading to $\mu_{GC0.5}(\boldsymbol{x},\boldsymbol{y})$ (see formula 4).

A. The labeling

Once the covariance has been computed for each segment, some distance measures are used in order to find the nearest reference for each segment (in a 24-dimensional space), as shown in the figure 1-a.

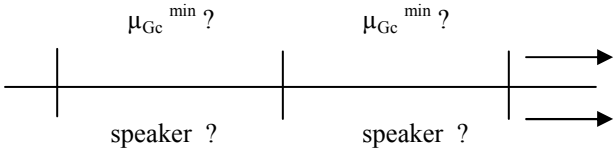

Fig. 1-a. Computation of the minimal distance.

Once the minimal distance between the segment features and the reference features (e.g. corresponding to speaker L_j) is found, the segment is labeled by the identity of this reference (speaker L_j), as shown in the

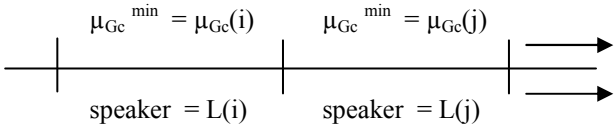

Fig. 1-b. Labeling.

$$\mu_{GC0.5}(\boldsymbol{x},\boldsymbol{y}) = \frac{\mu_{GC}(\boldsymbol{x},\boldsymbol{y}) + \mu_{GC}(\boldsymbol{y},\boldsymbol{x})}{2} \quad (2)$$

Another possibility for symmetrizing this measure is given in formula (5):

$$\mu_{GC\beta}(\boldsymbol{x},\boldsymbol{y}) = \frac{M\mu_{GC}(\boldsymbol{x},\boldsymbol{y}) + N\mu_{GC}(\boldsymbol{y},\boldsymbol{x})}{M+N} \quad (3)$$

This procedure of symmetrization can improve the classification performance, compared to both asymmetric terms taken individually.

4 SPEAKER DETECTION AND TRACKING

Speaker tracking is the process of following who says what in an audio stream [7]. Speaker tracking in our case is based on speaker detection: speaker verification techniques.

A. Interlaced segmentation

In our application, we divide the speech signal into two groups of uniform segments, in which each segment has a length of 2 seconds. The second segment group is delayed from the first one by a delay of 1 second, i.e. the segments are overlapped by 50%, as shown in figure 1-a. These two groups of segments, called respectively the odd sequence and the even sequence, form the interlaced segmentation.

In this survey we suppose that we have a total of 2n+1 numbered segments, in the speech signal, representing an odd sequence (1, 3, 5, 7, …, 2n+1) and an even sequence (2, 4, 6, 8, …, 2n).

Each segment is analyzed as followed: the speech signal is decomposed in frames of 512 samples (32 ms) at a frame rate of 256 samples (16 ms). The signal is not pre-emphasized. For each frame, a Fast Fourier Transform is computed and provides 256 square module values representing the short term power spectrum in the 0-8 kHz band. This Fourier power spectrum is then used to compute 24 filter bank coefficients. Thus, each segment is decomposed into several stationary frames (with 24 Mel-bank energy coefficients by frame) in order to compute its covariance.

figure 1-b. Thus, this process continues until the last segment of the speech file.

Finally, we obtain two labeling sequences corresponding to an even labeling and to an odd labeling, as shown in figure 2.

C. Interlaced speech indexing (ISI)

The ISI algorithm is a new technique in which there are two segmentations (one displaced from the other) and a logical scheme is used to find the best speaker labels, by combining the two segmentation sequences. See figure 2.

Having two different indexing sequences, we try to give a reasonable labeling compromise between the two previous labeling sequences. Thus, we divide each segment into two other similar segments (of 1 second each), called sub-segments, so that we obtain "2n" even labels (denoted by $L'^{1/2'}_{even}$) for the even sub-segments and "2n+2" odd labels (denoted by $L'^{1/2'}_{odd}$) for the odd sub-segments. Herein, $L'^{1/2'}_{even}$ and $L'^{1/2'}_{odd}$ are called sub-labels.

Our intuition would be that the even sub-label and the odd sub-label at the same sub-segment should be the same, therefore we must compare $L'^{1/2'}_{even}(j)$ with $L'^{1/2'}_{odd}(j)$ for each sub-segment j (for j=2, 3, … 2n+1). Herein, two cases are possible:

- **if** $L'^{1/2'}_{even}(j) = L'^{1/2'}_{odd}(j)$ **then** the label is correct:

$$\textbf{new label = correct label = } \mathbf{L'^{1/2'}(j) = L'^{1/2'}_{even}(j) = L'^{1/2'}_{odd}(j)} \quad (6)$$

where $L'^{1/2'}$ represents a sub-label.

- **if** $L'^{1/2'}_{even}(j) \neq L'^{1/2'}_{odd}(j)$ **then** the label is confused:

$$\textbf{new label = } \mathbf{L'^{1/2'}(j) = Cf} \quad (7)$$

where Cf means a confusion in the labeling.

In case of confusion (new label = Cf), we derive a new correction algorithm called "ISI correction".

Algorithm of ISI correction:

In case of confusion, we divide the corresponding sub-segments (of 1 s) into two other sub-segments of 0.5 second each, called micro-segments. Theirs labels, called micro-labels, are denoted by $L'^{1/4'}$.

The correction algorithm is then given by:

- **if** $\{ L'^{1/4'}(j) = Cf$ **and** $L'^{1/4'}(j+1) = Cf$ **and** $L'^{1/4'}(j-1) \neq Cf \}$

then $\mathbf{L'^{1/4'}(j) = L'^{1/4'}(j-1)}$ (8)

this is called a left correction (see the micro-segment j_0 in figure 2),

- **if** $\{ L'^{1/4'}(j) = Cf$ **and** $L'^{1/4'}(j-1) = Cf$ **and** $L'^{1/4'}(j+1) \neq Cf \}$

then $\mathbf{L'^{1/4'}(j) = L'^{1/4'}(j+1)}$ (9)

this is called a right correction (see the micro-segment j_1 in figure 2).

$L'^{1/4'}$ denotes a micro-label for a micro-segment of 0.5s.

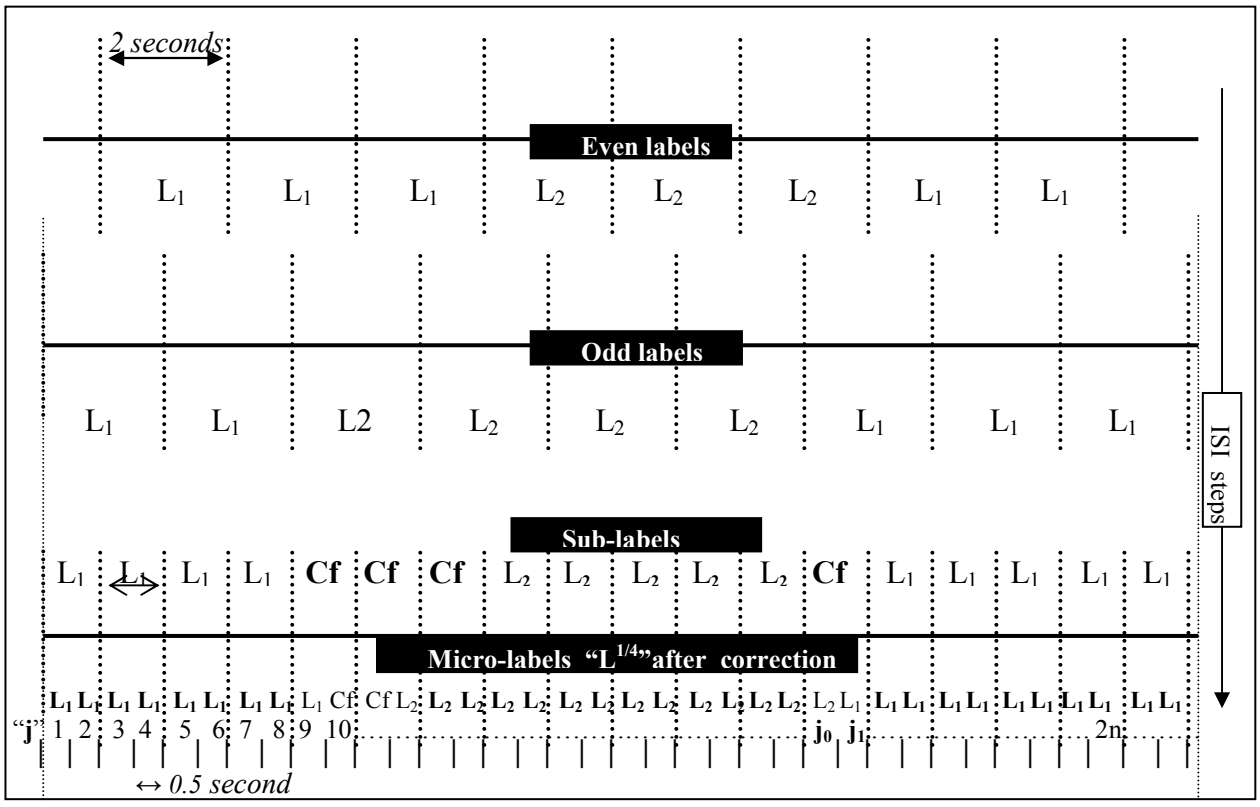

Fig. 2. The steps of the ISI algorithm with 1 iteration. L_j represents the speaker "j" and Cf means Confusion.

The ISI correction can be used several times (several iterations) to refine progressively the indexing accuracy. In our application we have used this algorithm with 2 and 4 iterations.

The experiments indicate that the ISI correction permits to find the best labeling decision, from the two interlaced labeling sequences, in reducing efficiently the tracking error. Moreover, the segmentation resolution (resolution of L(j)), which was 2 seconds, is reduced to only 0.5 second (resolution of $L^{1/4}(j)$), thus the performance brought by the ISI technique is observable both in the tracking accuracy and in the segmentation resolution.

V. Results and discussions

A. Experiments on TIMIT

The first test database consists of several utterances from TIMIT [11] uttered by different speakers and concatenated into speech files (the size of a speech file varies from 30 to 130 s), so that each speech file contains several sequences of utterances from different speakers (2, 3, 5 or 10 speakers per file) and with several speaker transitions per file. In order to investigate the robustness of our method, one part of the database is mixed with noise and music.

Thus, the global database represents 24 speech files of clean speech, 144 speech files of corrupted speech and 24 speech files of music-speech concatenation.

In this section we are interested in the different results obtained during the tracking tests on TIMIT database. All the results are summarized in table 1.

In table 1, we note that the tracking error increases if the number of speaker increases too. For example, in case of clean speech, the error is only 5.3% for 2 speakers and it is 7.3% for 3 speakers. Concerning the different noises added in this experiment, we see that human noise (e.g. cough, sneeze, "Euuh", "Heumm") do not disturb significantly the speaker tracking (the degradation is about 4% at 12dB) which implies that this type of noise may not disturb the tracking, considerably.

On another hand, background noise and office noise cause a high degradation of the tracking rate.

Concerning the music insertion, the results show that the presence of music does not degrade significantly the tracking performance.

Overall, the results obtained on this database are encouraging: the tracking error is low (5% for 2 speakers) without any degradation if music is inserted.

B. Experiments on Hub4 Broadcast News

The other speech data used in the experiments are extracted from the HUB-4 1996-Broadcast-News and consists of natural news recorded from the CNN broadcasting (news and interviews for about 30 minutes) which correspond to 19 different speakers. We test different measures and different correction algorithms in order to compare them strictly and objectively (see figures 3 and 4).

TABLE I: Tracking error for discussions between several speakers (2, 3, 5 or 10 speakers).

		Tracking error (%) for discussions between:			
		2 speakers	3 speakers	5 speakers	10 speakers
Clean speech	With silence detection	7,2	8,1	7,9	10,3
	Without silence detection	5,3	7,3	5,9	8,0
Music + speech	Without silence detection	4,8	6,6	7,5	9,1
Corrupted speech at 12 dB	Background noise	26,0	55,7	53,7	67,2
	Office noise	19,9	24,3	57,6	66,1
	Human noise	9,1	7,9	23,0	19,9
Corrupted speech at 6 dB	Background noise	32,8	58,4	64,7	79,1
	Office noise	28,1	37,7	63,4	70,6
	Human noise	11,8	12,9	15,5	24,3

Figure 3 represents the tracking error, on Hub4, obtained with different measures and for different segment durations. The different measures are described in section 3: μ_{Gc1} is a non-symmetric measure, $\mu_{Gc0.5}$ and $\mu_{Gc\beta}$ are symmetric (see formulas (3), (4) and (5).

Here, we note that the best measure is the $\mu_{Gc\beta}$ giving the best tracking performance. For example, if the segment duration is 3 second, μ_{Gc1} gives an error of 10.4%, $\mu_{Gc0.5}$ gives an error of 8.6% and $\mu_{Gc\beta}$ gives the smallest error, namely 7.7%.

In another way, this figure represents the tracking error, on Hub4, obtained with and without the use of ISI correction and for different segment durations, in order to get a global comparison. Here we note that the tracking error obtained after ISI correction is lower than that obtained without ISI correction. For example, if the segment duration is 3 seconds, the error of tracking without ISI correction is about 9% but it decreases to 7.7% when an ISI correction with two iterations is applied and decreases to 7.6% when an ISI correction with four iterations is applied.

Finally, another comparison between the different tested durations shows that the smallest tracking error is obtained for the segment duration of 3 s.

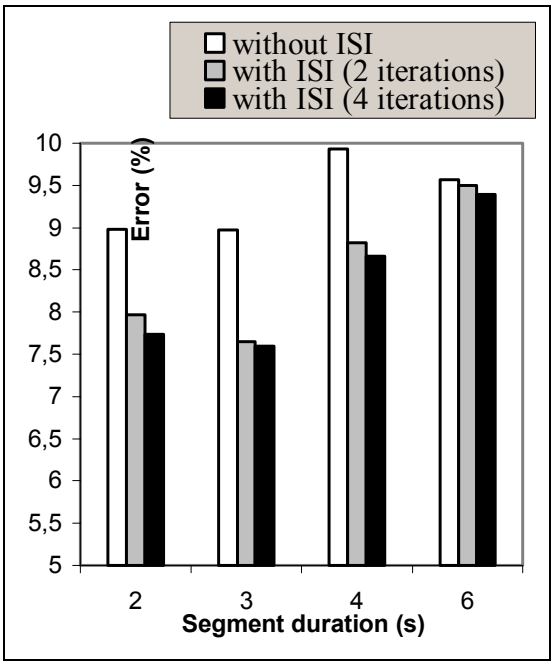

Fig. 3. Tracking error for a tracking with or without ISI correction.

VI. CONCLUSION

We have conceived a new method for automatic speaker tracking and detection, called "ISI", using an interlaced equidistant segmentation. Experiments indicate that the association between the SOSM method and the ISI technique is used efficiently: Although the SOSM needs a speech duration of at least 2 seconds, which means that the segmentation resolution is about 2 seconds, the association SOSM-ISI allows a finer resolution by reducing the segmentation error to only 0.5 second. Furthermore, this new approach improves considerably the tracking accuracy by correcting the confusion errors. Moreover, when we increase the number of corrections in the ISI algorithm (i.e. number of iterations) the tracking error decreases continuously.

For the evaluation, two databases (TIMIT and Hub4) are used:

- In the first evaluation, the test database consists of several utterances from TIMIT [11] uttered by different speakers and concatenated into speech files (the size of a speech file varies from 30 to 130 s). In order to investigate the robustness of our method, one part of the database is mixed with noise and music.
- In the second evaluation, the speech data are extracted from the 1996 Broadcast News and consists of natural news recorded from the CNN broadcasting (news and interviews for about 30 minutes).

The best performance is obtained, on TIMIT, with a tracking score of about 95% (percentage of correctly labeled segments), if no noise is mixed with the speech signal. When noise is mixed, the tracking score decreases with the SNR, but the experiments show that human noise does not disturb significantly the speaker tracking. Moreover, when a pure music is inserted inside the speech signal (by concatenation) the tracking score remains unchanged. Also, the results suggest that the tracking error increases when the number of speakers increases (which is obvious) and it decreases when the speakers have different sex, because their features are very different.

The experiments done on Hub4 Broadcast News indicate that the best statistical measure is the $\mu_{Gc\beta}$ (this result was also reported by Bimbot in [6]) and that the best segment duration for the tracking is 3 seconds.

Compared to previous results (section 2), our method provides interesting results. Although it is difficult to compare objectively the performances of all the existing methods, we believe that our new method (ISI) represents a good speaker tracking approach, since it is easy to implement, simple to reproduce, inexpensive in computation and provides a high tracking performance.

REFERENCES

[1] H. Gish, "Robust discrimination in automatic speaker identification". IEEE International Conference on Acoustics Speech and Signal Processing. April 90, New Mexico, 289-292.

[2] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan: "Second-Order Statistical measures for text-independent Broadcaster Identification". Speech Communication, Volume 17, Number 1-2, August 1995, 177-192.

[3] J.F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin and C.J. Wellekens, "A speaker tracking system based on speaker turn detection for NIST evaluation". IEEE International Conference on Acoustics Speech and Signal Processing. Istanbul, june 2000.

[4] P. Delacourt , "Indexing de données audio: segmentation et regroupement par locuteurs". PhD thesis , Ecole Normale Supérieure des Télécommunications, 2000, Paris France.

[5] D. Liu, and F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing". Eurospeech, 1999. Vol. 3, 1031-1034.

[6] D.A. Reynolds, E. Singer, B.A. Carlson, G.C. O'Leary, J.J. McLaughlin, and M.A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics". International Conference on Spoken Language Processing, 1998. Vol. 7, 3193-3196.

[7] W. Fisher, V. Zue, J. Bernstein, and D. Pallet, "An acoustic-phonetic database", 1986, JASA, suppl. A, Vol. 81(S92).