Speaker based segmentation on broadcast news-on the use of ISI technique

S. Ouamour[1], M. Guerti[2] and H. Sayoud[1]
[1] USTHB, Electronics Institute, BP 32 Bab Ezzouar, Alger, Algeria
[2] ENP, Hacen Badi, El-Harrach Alger, Algeria

Abstract

In this paper we propose a new segmentation technique called ISI or "Interlaced Speech Indexing", developed and implemented for the task of broadcast news indexing. 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 and then to his talk. Our segmentation procedure is based on an interlaced equidistant segmentation (IES) associated with our new ISI algorithm. This approach uses a speaker identification method based on Second Order Statistical Measures. As SOSM measures, we choose the "µGc" one, which is based on the covariance matrix. However, 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 this association provides a high resolution and a high tracking performance: the indexing score (percentage of correctly labelled segments) is 95% on TIMIT database and 92.4% on Hub4 Broadcast news 96 database.

Introduction

Speaker tracking 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), there are 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 (keyword, 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 this task 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 indexing can be seen, herein, as a speaker verification task applied locally along a document containing multiple (and unknown) interventions of vari-

Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, 28-30 August 2006, Athens, Greece.

ous 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. This algorithm is easy to implement, simple and efficient since it significantly improves the results.

Speaker detection and tracking

Speaker tracking is the process of following who says what in an audio stream (Delacourt 2000, Bonastre 2000). Our speaker identification method is based on mono-Gaussian models and uses some measures of similarity called Second Order Statistical Measures (Gish 1990, Bimbot 1995). In our experiments we used the µGc measure (based on the covariance matrix).

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%. These two groups of segments, called respectively the odd sequence and the even sequence, form the interlaced segmentation.

B. Labeling

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

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). Thus, this process continues until the last segment of the speech file. Finally, we obtain two labeling sequences corresponding to an even labeling and an odd labeling, as shown in figure 1.

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.

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. Herein, two cases are possible:

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

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

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

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

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

In case of confusion, we derive a new 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 $\mathbf{L^{'1/4'}(j) = L^{'1/4'}(j-1)}$ (3)
  this is called a left correction (see the micro-segment j_0 in figure 1),

- **if** { $L^{'1/4'}(j) = Cf$ **and** $L^{'1/4'}(j-1) = Cf$ **and** $L^{'1/4'}(j+1) \neq Cf$ }
  then $\mathbf{L^{'1/4'}(j) = L^{'1/4'}(j+1)}$ (4)
  this is called a right correction (see the micro-segment j_1 in figure 1).

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

Results and discussions

The first test database consists of several utterances from TIMIT uttered by different speakers and concatenated into speech files.

Table 1: Tracking error for discussions between several 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

Each speech file contains several sequences of utterances from different speakers 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. In table 1, we note that the tracking error increases if the number of speakers 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 do not disturb significantly the speaker tracking (degradation of 4% at 12dB) which implies that this type of noise may not disturb the tracking, considerably.
The other speech data used in the experiments are extracted from the *HUB-4 1996-Broadcast-News* and consists of natural news.

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.

Moreover, we notice that the best tracking is got for segments duration of 3s.

Conclusion

Experiments done on corrupted speech and on *Hub4 Broadcast News* indicate that the ISI technique improve both the indexing precision and the segmentation resolution. Furthermore, they show that the best segment duration for speech segmentation is 3 seconds.

In general, compared to previous works, this method gives interesting results. Although it is difficult to compare objectively the performances of all the existing methods, we believe that this technique represents a good speaker indexing approach, since it is easy to implement, inexpensive in computation and provides good performances.

References

Bimbot F. et al. 1995. Second-Order Statistical measures for text-independent Broadcaster Identification. Speech Communication, 17, 177-192.

Bonastre J.F. et al. 2000. A speaker tracking system based on speaker turn detection for NIST evaluation. IEEE ICASSP, Istanbul, june 2000.

Delacourt P. et al. 2000. DISTBIC: a speaker-based segmentation for audio data indexing, Speech Communication, 32, Issue 1-2.

Gish H. 1990. Robust discrimination in automatic speaker identification. IEEE Inter. Conference on Acoustics Speech and Signal Processing. April 90, New Mexico, 289-292.

Liu D., and Kubala F. 1999, “Fast speaker change detection for broadcast news transcription and indexing”. Eurospeech, 1999. Vol. 3, 1031-1034.

Reynolds D.A. et al. 1998, “Blind clustering of speech utterances based on speaker and language characteristics”. ICSLP, 1998. Vol. 7, 3193-3196.