Automatic Speaker Indexing in Corrupted Speech.

H. SAYOUD*, S. OUAMOUR**, M. BOUDRAA
*SSTR Laboratory - No: 104 Djenane-Mabrouk, Badjarah, Alger, Algeria.
Email: * sayoud@ifrance.com , ** ouamour@ifrance.com

ABSTRACT **--- Speaker indexing can broadly be divided into two problems: Locating the points of speaker change (Segmentation) and Identifying the speaker in each segment (Labeling). An important obstacle, in the speaker tracking, is the corruption of the speech signal during its recording or in a telephonic conversation.**
In this paper, we are interested in the corruption of the speech signal by the most probable noises during audio-visual recording and the mixture of the speech signal with music, in order to test the robustness of our speaker tracking method. For this purpose, we choose the SOSM method (Second Order Statistical Measures), applied for segments of 2 seconds duration with an overlapping of 50%. The speaker indexing becomes very difficult if the recordings are made in a noisy environment or if music is mixed with the speech.
The evaluation of our method is done in TIMIT, and each discussion consists on sequences of speech signals uttered by 2 different speakers, concatenated into one speech file (the speakers are arbitrarily chosen from a population of 37 different speakers). So, each speech file contains several speaker transitions by file.
In a second step, we have corrupted the database by three types of noise, namely: the office noise, the human noise and the background noise. Moreover, we have inserted music inside the discussion signals, for example, at the beginning, at the middle and at the end of the discussion.
The results got are severely discussed according to each case: clean environment, noised environment, presence of music, etc. As an example, the error rate of the tracking varies from 5%, in a clean environment, to 34%, in a noised environment (+6 dB).
Moreover, we remark that the error rate increases when the SNR decreases. Concerning the music, we remark that the speaker indexing is not perturbed by the concatenation of the music sequences, which is interesting in the case of the musical advertisement.

I. INTRODUCTION

The speaker indexation has many applications. We can give some examples of applications, like the indexation of the audio stream recorded from a radio (in order to track a speaker) or like the automatic speaker tracking by camera during teleconferences or seminars (without human help).
For the last example, some systems based on microphone arrays do exist; however, they are limited due to certain restrictions they place. Fortunately, recent progress in signal processing technologies is making it feasible to start automating the audiovisual supervision for capturing seminars. Our research focuses on systems designed for the audiovisual supervision of conferences (tracking systems), using speaker recognition methods based on statistical measures. The goal of this work is to investigate the development of an affordable and portable speaker indexing system capable of locating and tracking speakers in noisy environment [**Ros98]**.

II. THE SOSM-BASED METHOD

This method, for speaker identification, is based on mono-Gaussian statistical models. It is used in order to recognize the speaker identity at each segment of the speech signal.

A brief description is given bellow.

Let $\{x_t\}_{1\le t\le M}$ be a sequence of M vectors resulting from the p-dimensional acoustic analysis of a speech signal uttered by speaker **X**. These vectors are summarized by the mean vector $\bar{x}$ and the covariance matrix X:

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \qquad (1)$$

and

$$X = \frac{1}{M}\sum_{t=1}^{M}(x_t - \bar{x})(x_t - \bar{x})^T \qquad (2)$$

Similarly, for a speech signal uttered by speaker **Y**, a sequence of N vectors $\{y_t\}_{1\le t\le M}$ can be extracted.

By supposing that all acoustic vectors extracted from the speech signal uttered by speaker **X** are distributed like a Gaussian function, the likelihood of a single vector y_t uttered by speaker **Y** is

$$G(y_t / X) = \frac{1}{(2\pi)^{p/2}(\det X)^{1/2}} e^{(1/2)(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})} \quad (3)$$

If we assume that all vectors y_t are independent observations, the average log-likelihood of $\{y_t\}_{1\le t\le M}$ can be written as

$$\bar{L}_x(y_1^N) = \frac{1}{N}\log G(y_1 \ldots y_N|\mathbf{X}) = \frac{1}{N}\sum_{t=1}^{N}\log G(y_t|\mathbf{X}) \quad (4)$$

We also define the minus-log-likelihood $\mu(\mathbf{X}, y_t)$ which is equivalent to similarity measure between vector y_t (uttered by **Y**) and the model of speaker **X**, so that

$$Arg\max_{X} G(y_t / X) = Arg\min_{X} \mu(X, y_t) \quad (5)$$

We have then:

$$\mu(\mathbf{X}, y_t) = -\log G(y_t / \mathbf{X}) \quad (6)$$

The similarity measure between test utterance $\{y_t\}_{1\le t\le M}$ of speaker **Y** and the model of speaker **X** is then

$$\mu(X,Y) = \mu(X, y_1^N) = \tfrac{1}{N}\sum_{t=1}^{N}\mu(X, y_t) \quad (7)$$

$$= -\bar{L}_X(y_1^N) \quad (8)$$

After simplifications, we obtain

$$\mu(X,Y) = \frac{1}{P}\left[-\log(\frac{\det(Y)}{\det(X)}) + tr(YX^{-1}) + (\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\right] - 1 \quad (9)$$

This measure is equivalent to the standard Gaussian likelihood measure (asymmetric μ_G) defined in [Bim95].

A variant of this measure called μ_{Gc} is deduced from the previous one in supposing that

$\bar{y} = \bar{x}$ (i.e. the inter-speaker variability of the mean vector is negligible).

Thus the new formula becomes:

$$\mu_{GC}(X,Y) = \frac{1}{P}\left[-\log(\frac{\det(Y)}{\det(X)}) + tr(YX^{-1})\right] - 1 \quad (10)$$

All measures reviewed in this section have the common property of being non-symmetric.

In other words, the roles played by the training data and by the test data are not interchangeable.

However, our intuition would be that a similarity measure should be symmetric.

A simple possibility for symmetrizing this measure $\mu_{GC}(Y,X)$ is to construct the average between the measure and its dual term:

$$\mu_{GC0.5}(X,Y) = \frac{\mu_{GC}(X,Y) + \mu_{GC}(Y,X)}{2} \quad (11)$$

This procedure of symmetrization can improve the classification performance, compared to both asymmetric terms taken individually. This measure $\mu_{GC0.5}$ will be used in the experiments described in this paper.

III. Speaker Indexing

A. Description of the problem

Speaker Indexing is the process of following who says what in an audio stream **[Bon-ICASSP00, Bon00, Gau98, Liu99, Nis98, Nis99, Rey98]**.

Speaker indexing has many applications, for example in the political broadcasting a correct behaviour imposes on candidates that campaign to the Chamber of Representatives or for President, to use equal time for their public TV or radio addresses. The control of the use of broadcasting media is checked manually (in France by the *"Conseil Supérieur de l'Audiovisuel"*): automatic recording of the debates could ease this task. **[Del2000]**.

In our application, the tracking can be divided into two problems:

- Locating the points of speaker change (Segmentation).
- Identifying the speaker in each segment (Labeling).

Segmentation can be thought of as labeling on a very fine scale. For example consider the case of having two distinct segments. Suppose you can accurately determine whether they originate from the same speaker or different speakers. This means the labeling problem has been solved. A simple segmentation can be achieved by regularly generating segments throughout the audio and then joining together the adjacent segments which originate from the same speaker. This has a particular advantage in that it works very quickly, but in the other hand the resolution will be coarse. Suppose it takes 2 seconds worth of speech, to produce the information for a segment which allows it to be identified. Then the point of speaker change will be uncertain to within roughly 2s. This problem can be overcome to some extent by using an interlaced indexing algorithm (to be published) which reduces the indexation resolution to only 0.5s.

The labeling problem reduces to finding a representation of each segment which captures the information about the speaker, whilst, if possible, minimizing the intra-speaker variation. These representations can then be compared to each other to ascertain which ones are most similar and hence determine which speakers uttered which segments **[Ros98]**.

Finding such a representation is a difficult problem. For example, if the speech is coded in PLP parameters, taking the mean vector over a small segment may retain some speaker-specific information (such as gender), but it will also be highly dependent on which phoneme was being uttered at that time. One method of reducing intra-speaker variation, which has already been used in problems similar to this one is using the covariance of the data, as the SOSM method **[Bim95, Bim96, Bon97]**, over a reasonably sized segment (at least 2 seconds of speech). This method is text-independent i.e. it does not require a transcription of what was said, but instead effectively averages out the phoneme variation over the segment.
Speaker clustering is concerned more with the improvement of speaker-adaptive recognition systems **[Bon-ICASSP00, Bon00, Gau98, Liu99, Nis98, Nis99, Rey98]**. Segments are clustered into groups which are in some sense more similar to members of their own group than those of the other groups. The ideal case would be if every cluster represented a different speaker, but this is obviously dependent on the number of final clusters and the number of speakers in the soundtrack (which is not necessarily known in advance).

B. Segmentation

In our application, we divide each speech signal into two groups of equidistant segments and each segment has a length of 2 seconds.
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 will be decomposed into several stationary frames (with 24 Mel-bank energy coefficients by frame) in order to compute its covariance.

C. Silence detection

The principle of segmentation with respect to speakers based on silence detection relies on the assumption, not always verified, that utterances of different people are separated by significant silences. To detect inter-speaker silences, Nishida and Ariki **[Nis98]** use the average power of the speech signal. If the power value is below a given threshold, then the signal is identified as silence. The authors do not give any details about how they choose the threshold. It may be tuned for each recording.
In our project, we use the silence detection in order to refine the speaker tracking, but we do not detect the silence in all the tests.

D. The labeling

Once the covariance has been computed for each segment, some measures of distance must be used to calculate the closeness of the reference speakers in each segment (in a 24-dimensional space), as shown in the figure below.
Once, the minimal distance between the segment and the reference model (suppose that it corresponds to the speaker L_j) is found, then the segment is labeled by the identity of this speaker L_j.

Then, we continue this process until the last segment in the speech file. Finally, we obtain two label sequences corresponding to the two segmentation sequences, which are used by our new post-processing algorithm (to be published).

IV. SPEECH DATABASE

A. Description of the Database

The test database consists of several utterances from TIMIT **[Fis86]** uttered by different speakers, concatenated into speech files, so that each speech file will contain several sequences of utterances from different speakers. Thus each speech file can contain two, three, five or ten utterances from different speakers, with several speaker transitions per file. The duration of a speech file is between 30 and 130 seconds. In order to complicate the tests, one part of the database is mixed with different noises and different types of music **[Mon98]**. The global database represents 24 speech files of clean speech, 144 speech files of corrupted speech and 24 speech files containing an association of music and speech.

B. Corrupted database

We have corrupted the different speech files by 3 types of noise, usually frequent in seminars and teleconferencing. They are:

- The human noise, corresponding to the different sounds produced by the human being, as the cough, the sneeze or the brief sounds like “Euuh”, “Heumm”, etc.
- The office noise, like sounds produced by moving chairs or ashtrays or like sounds produced by the paper rustling.
- The background noise, caused by the electronic devices or the recording equipments.

Thus, the speech signals are corrupted by these three types of noise at +12dB and at +6 dB.

C. Speech-music database

Music simulates the musical advertisements recorded during the recording of a conference or an interview. So we choose a variety of 10 types of music (each

music sequence has a length of 10 seconds) like classical music, jazz, rock, etc. In our application, music is concatenated with speech at the beginning, at the end or inside the speech file.

V. RESULTS AND DISCUSSION

In this section we are interested in the different results got during the tests in TIMIT. All these results are summarized in tables 1 and 2.

Table 1 shows the different error rates obtained during the automatic tracking of 2 speakers who are speaking in different conditions, i.e. with clean speech, with corrupted speech, with music and without music.

In this table, we notice that the best performance is obtained for an error rate of approximately 5%, namely it will be impossible to give an error rate lower than 5%, if the tracking method (as described in this paper) is SOSM **[Bim95, Bim96, Bon97]**. We think that an error rate of 5% is sufficient to track efficiently the speakers. However, in this experiment we have used a high quality speech (TIMIT), but in reality the speech can be corrupted or distorted, so the tracking error should be lower in such condition.

Concerning the different noises added in this project (see table 1), we notice that human noise (cough, sneeze, "Euuh", "Heumm", …) do not disturb, significantly, the speaker tracking (degradation of about 4% at 12dB) which implies that this type of noise will not disturb the audiovisual tracking, considerably.

In the other hand, background noise and office noise (sounds produced by moving chairs and ashtrays or produced by paper rustling) cause a high degradation of the tracking rate. So, the conference (or the teleconference) organizers must provide high-quality recording equipment and must demand the speakers to avoid moving objects on the desk, if this moving can cause noises during the recording. We also notice, in the same table, that the error will increase if the number of speakers increases. For example, in case of clean speech, the error is only 5.3% for 2 speakers.

Table 2 is the same as the first table, except that it presents a particular classification based on the sex of speakers. So we can have 3 cases of discussions:

1- discussions between two female speakers (in the 3rd column),
2- discussions between two male speakers (in the 4th column),
3- discussions between a female speaker and a male speaker (in the 5th column).

Here, we notice that the least error rate is obtained when speakers sexes are different (this is observable in the case of clean speech).

Consequently, the tracking will be better if speakers have different sexes (in a 2-speakers discussion). So it will be profitable, for example, to choose a female journalist if the interviewed minister is a man.

More over, the error rate remains unchanged even if music is mixed with speech. (table 1 and 2).

Concerning the music insertion, tables 1 and 2 show that we do not note any degradation in the tracking score. Since the presence of this music doesn't degrade the tracking performance, we can authorize the insertion of music (a pure music but not a song) inside multi-speaker discussions (at the beginning, the middle or at the end of the discussion) without any hesitation.

Globally for this database, we think that these results are encouraging, because our system permits to track speakers with a low tracking error (5% for 2 speakers) and with a low segmentation error (delay of only 0.5s), without any degradation if music is inserted.

Table 1 Tracking error for discussions between 2 speakers.

		Indexing error (%) for discussions between 2 speakers
Clean speech	With silence detection	7,15
	Without silence detection	5,3
Music + speech	Without silence detection	4,84
Corrupted speech at 12 dB	Background noise	25,95
	Office noise	19,89
	Human noise	9,14
Corrupted speech at 6 dB	Background noise	32,84
	Office noise	28,05
	Human noise	11,84

Table 2 Tracking error according to the sex of speakers.

		Indexing error (%) for discussions between 2 speakers:		
		female speaker + female speaker	**mal speaker + mal speaker**	**female speaker + mal speaker**
Clean speech	With silence detection	8,6	7,17	5,67
	Without silence detection	5,93	5,59	4,39
Music + speech	Without silence detection	4,59	4,76	5,17
Corrupted speech at 12 dB	Background noise	31,76	26,75	19,33
	Office noise	17,2	20,03	22,44
	Human noise	11,32	10,84	5,61
Corrupted speech at 6 dB	Background noise	39,71	34,49	24,31
	Office noise	20,53	31,14	32,47
	Human noise	12,89	13,99	8,63

VI. CONCLUSION

We develop a new method for automatic speaker tracking, using a statistical measure called SOSM. In our evaluation, the speech signals consist of several utterances from different speakers (extracted from TIMIT) and concatenated into speech files, so that each speech file will contain several utterances from different speakers. Thus each speech file can contain different speakers with several speaker transitions per file. In order to simulate a noisy environment, speech is mixed with different types of noise and is concatenated with different music sequences.

The experimental results show that the best performance is obtained with an indexing score of about 95% (percentage of correctly labeled segments), if no noise is mixed with the speech signal.

When noise is mixed, the indexing score decreases with the SNR, but the experiments show that human noise doesn't disturb significantly the speaker tracking. More ever, when a pure music is inserted inside the speech signal (by concatenation) the indexing score remains unchanged. This proves that we can insert music inside multi-speaker discussions (at the beginning, the middle or at the end of the discussions).

A special classification of the results shows that the tracking error decreases when the speakers have different sexes. Consequently, the tracking will be more efficient if the sexes of speakers are different (in a 2-speakers discussion, for example).

The experiments show that the µGc[0.5] measure is very effective in the speaker indexing, because it permits to identify accurately the different speakers, even if the speech is corrupted. Then we can prove that the SOSM technique is efficient and robust in the speaker indexing.

REFERENCES

[Bau94] Baumberg, A. & Hogg, D. An efficient method for contour tracking using active shape models. Technical Report *TR 94.11*, University of Leeds.

[Bei86] Beigi, H.S.M., Maes, S. Speaker, channel and environment change detection. In: World Congress of Automation, WCA 98.

[Bim95] F. Bimbot, I. Magrin-Chagnolleau, et L. Mathan "Second-Order Statistical measures for text-independent Broadcaster Identification". Speech Communication, Volume. 17, Number, 1-2, August 1995, pp. 177-192.

[Bim96] I. Magrin-Chagnolleau, J. Wilke, F. Bimbot "A Further Investigation on AR-Vector Models for Text-Independent Speaker Identification". ICASSP, pp 401-404, May 7-10 1996.

[Bon97] J.F. Bonastre et L. Besacier "Traitement Indépendant de Sous-bandes Fréquentielles par des méthodes Statistiques du Second Ordre pour la Reconnaissance du Locuteur". Actes du 4ème Congrès Français d'Acoustique, pp 357-360, Marseille 14-18 April 1997.

[Bon-ICASSP00] Bonastre, J.F., Delacourt, P., Fredouille, C., Merlin, T., Wellekens, C.J. 2000. A speaker tracking system based on speaker turn detection for NIST evaluation. In: IEEE International Conference on Acoustics Speech and Signal Processing.

[Bon-JEP00] J.F. Bonastre et Al, " Modèle de Markov évolutif pour les tâches de suivi de locuteurs ". JEP'2000, Aussois, France, pp 69-72, 19-23 juin 2000.

[Chen98] Chen, S.S., Gopalakrishnan, P.S. Speaker environment and channel change detection and clustering via the Bayesian Information Criterion. In: DARPA Speech Recognition Workshop, DARPA-SRW 98.

[Cut98] Cutler, R. Cutler and Turk, M. View-based Interpretation of Real-time Optical Flow for Gesture Recognition. IEEE Automatic Face and Gesture Recognition, April 1998.

[Dau83] B.A. Dautrich, L.R. Rabiner and T.B. Martin "The effects of selected signal processing techniques on the performance of a filter bank based isolated word recognizer". Bell System Technical Journal, 1983.

[Del00] P. Delacourt, "Indexing de données audio: segmentation et regroupement par locuteurs". PhD thesis , Ecole Normale Supérieure des Télécommunications, Paris, France.

[Dod98] G. R. Doddington, "Speaker Recognition Evaluation Methodology. An Overview and Perspectives". RLA2C Avignon France, 20-23 April 1998, pp 60-66.

[Fis86] W. Fisher, V. Zue , J. Bernstein and D. Pallet, “An acoustic-phonetic database”. JASA, suppl. A, Vol. 81(S92) 1986.
[Gau98] Gauvain, J.-L., Lamel, L., Adda, G. Partitioning and transcription of broadcast news data. International Conference on Spoken Language Processing 4, pp 1335-1338.
[Gis91] Gish, H., Siu, M.-H., Rohlicek, R. Segregation of speakers for speech recognition and speaker identification.
In: IEEE International Conference on Acoustics Speech and Signal Processing. pp. 873-876. [Gis94] Gish, H., Schmidt, N. Text-independent speaker identification. In: IEEE Signal Processing Magazine, October, 18-32.
[Liu99] Liu, D., Kubala, F. Fast speaker change detection for broadcast news transcription and indexing. In: Eurospeech. Vol. 3, pp. 1031-1034.
[Liu00] Q. Liu, Y. Rui, A. Gupta and JJ. Cadiz, Automating Camera Management for Lecture Room Environments. Technical Report No: MSR-TR-2000-90, Microsoft Research, sept. 2000, Microsoft.
[Mon98] Montacie, C., Caraty, M.-J. A silence/noise/music/speech splitting algorithm. In: International Conference on Spoken Language Processing. Vol. 4, pp. 1579-1582.
[Muk99] Mukhopadhyay, S., & Smith, B. Passive Capture and Structuring of Lectures. Proc. Of ACM Multimedia’99, Orlando.
[Nis98] Nishida, M., Ariki, Y. Real time speaker indexing based on subspace method: applications to TV news articles and debate. In: International Conference on Spoken Language Processing. Vol. 4, pp. 1347-1350.
[Nis99] Nishida, M., Ariki, Y. Speaker indexing for news articles debates and drama in broadcasted TV programs. In: IEEE International Conference on Multimedia Computing and Systems. pp. 466-471.
[Rey98] Reynolds, D.A., Singer, E., Carlson, B.A., O'Leary, G.C., cLaughlin, J.J., Zissman, M.A. Blind clustering of speech utterances based on speaker and language characteristics. In: International Conference on Spoken Language Processing. Vol. 7, pp. 3193-3196.
[Par01] ParkerVision. http://www.polycom.com
[Ris89] Rissanen, J. Stochastic Complexity in Statistical Inquiry.
Series in Computer Science, Vol. 15. World Scientific, Singapore, Chapter 3.
[Ros98] Rosenberg, A.E., Magrin-Chagnolleau, I., Parthasarathy, S., Huang, Q., Speaker detection in broadcast speech databases. In: International Conference on Spoken Language Processing. Vol. 4, pp. 1339-1342.
[Sie97] Siegler, M.A., Jain, U., Raj, B., Stern, R.M. Automatic segmentation classification and clustering of broadcast news audio. In: DARPA Speech Recognition Workshop. pp. 97-99.
[Sti99] Stiefelhagen, R., Yang, J., & Waibel, A. Modeling focus of attention for meeting indexing. Proc. of ACM Multiemdia’99.
[Tri98] Tritschler, A. A segmentation-enabled speech recognition application using the BIC criterion. Master's thesis. Institut EURECOM, France.
[Woo97] Woodland, P.C., Gales, M.J.F., Pye, D., Young, S.J. The Development of the 1996 HTK broadcast news transcription system.
In: DARPA Speech Recognition Workshop. pp. 97-99.