Digital Signal Processing **10**, 172–197 (2000)
doi:10.1006/dspr.1999.0367, available online at http://www.idealibrary.com on IDEAL®

AMIRAL: A Block-Segmental Multirecognizer Architecture for Automatic Speaker Recognition

Corinne Fredouille, Jean-François Bonastre, and Teva Merlin

Laboratoire Informatique d'Avignon (LIA), Agroparc BP 1228, 84911 Avignon Cedex 9, France

E-mail: corinne.fredouille@lia.univ-avignon.fr, jean-francois.bonastre@lia.univ-avignon.fr, teva.merlin@lia.univ-avignon.fr

Fredouille, Corinne, Bonastre, Jean-François, and Merlin, Teva, AMIRAL: A Block-Segmental Multirecognizer Architecture for Automatic Speaker Recognition, *Digital Signal Processing* **10** (2000), 172–197.

In the wide domain of automatic speech recognition, extracting the relevant information carried by the speech signal is far from easy. Diversity, redundancy, and variability, the main characteristics of the speech signal, make this task particularly difficult. The work reported here presents a multirecognizer architecture designed to cope with this issue in the framework of Automatic Speaker Recognition. This architecture, based on various individual recognizers, exploits different classes of information conveyed by the speech signal. In this paper, two classes of information are investigated: information related to the frequency domain, and "dynamic" information. This multirecognizer architecture is coupled with a block-segmental approach applied on each classifier. The overall system allows us to emphasize the most informative temporal blocks and to discard the least informative ones or those corrupted by noise. The AMIRAL system developed by the LIA integrates both approaches and was tested during the NIST/NSA 1999 speaker recognition evaluations. The results of these experiments for the tasks of Speaker Verification ("One Speaker" and "Two Speakers") and Speaker Tracking are provided and discussed.

Key Words: Multirecognizer architecture; block-segmental approach; frequency domain; dynamic information; automatic speaker verification; MAP normalization.

1. INTRODUCTION

Speech, as a human faculty, is full of meaning and conveys information from various origins (linguistics, emotional state of the speaker, etc.). This richness as well as the complexity of the underlying communication process make its medium—the speech signal—extremely variable and sensitive to

1051-2004/00 $35.00

the environment. To ensure the robustness of the communication, a strong redundancy between the various classes of information is present in the speech signal. This phenomenon may also be observed within a single class of information in order to transmit a precise message—for instance, stress (prosodic information) may be expressed simultaneously by pitch variation, energy variation, and syllable extension.

In the wide domain of automatic speech processing, the variability as well as the redundancy of the speech signal has to be taken into account. This work proposes an original architecture designed to cope with these issues in the framework of Automatic Speaker Recognition.

To take advantage of the redundancy between different types of speaker-specific information, a multirecognizer approach was proposed in which all the recognizers worked in parallel [8]. This approach, described in Section 2, is completed by a segmentation and normalization process applied to each recognizer. The proposed method allows the most informative temporal blocks to be emphasized and the least relevant ones or those corrupted by noise to be discarded. This "block-segmental" technique is defined in Section 3. Section 4 presents the AMIRAL system (see Fig. 1), developed by the LIA, [1] which integrates both proposed approaches and details the experimental conditions used for their validation. Experiments conducted on different multirecognizer architectures in the framework of Automatic Speaker Verification are commented on in Section 5. These architectures involve multiple frequency subbands as well as several methods for the exploitation of dynamic information. The adaptation of the AMIRAL system to the speaker tracking task is presented in Section 6. Finally, Section 7 discusses the results obtained and proposes further investigation.

2. MULTIRECOGNIZER ARCHITECTURE

Among the huge amount of information conveyed by the speech signal multiple sources of speaker-specific information may be isolated, such as information carried by short-term and long-term spectra, prosodic information, and phonetical/articulatory information such as phoneme instantiation, coarticulatory phenomena, or formant trajectory. This nonexhaustive list illustrates the diversity of information categories used for speaker characterization. The methods and criteria used to exploit data may differ from one category of information to another. Therefore, these various sources of information require specific processing for speaker recognition tasks.

In this paper, the following classes of information are investigated: information related to the frequency bands of short-term spectra (Section 2.1) and the evolution of these spectra (Section 2.2). The speaker-specific information is emphasized thanks to a multirecognizer architecture in which each kind of information is associated with a unique recognizer and a specific processing.

[1] Laboratoire Informatique d'Avignon.

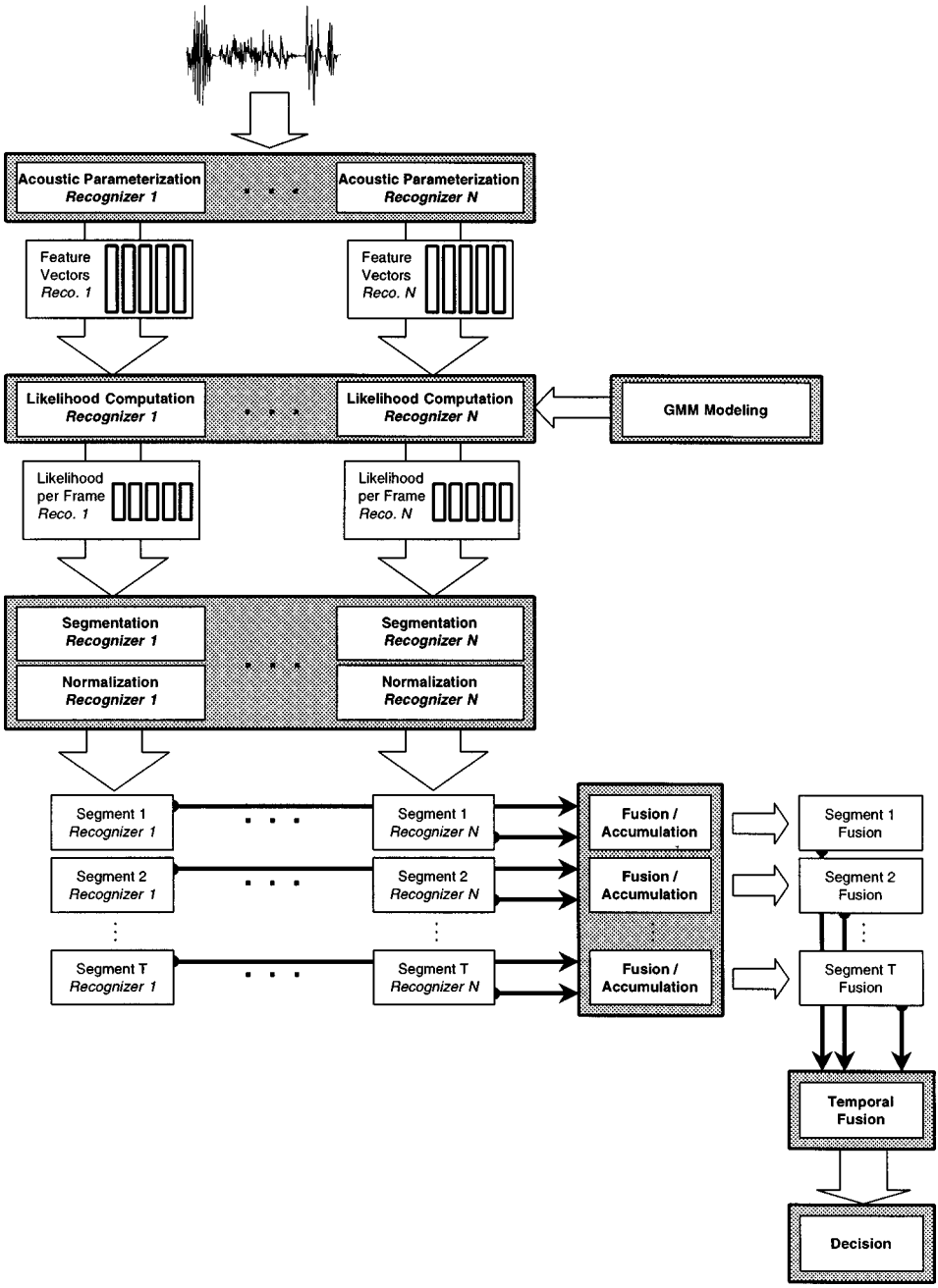

FIG. 1. The AMIRAL system. Details of the different modules integrated in the AMIRAL system.

2.1. Frequency Subbands

Splitting the spectral domain into individual subbands has been a widely used approach in automatic speech recognition [10, 15, 24] and Automatic Speaker Recognition (ASR) [4, 8]. Two main studies [2, 17] suggest that the

human-being recognition process is based on individual frequency subbands processed independently from each other. Besacier *et al.* [9] demonstrate that useful information for Automatic Speaker Identification (ASI) is mainly based on the low ($f < 500$ Hz) and high ($f > 2500$ Hz) frequency bands. [2] The above-mentioned studies show an improvement of performance in automatic speech recognition and ASI. Few studies report the use of frequency subbands for the task of Automatic Speaker Verification (ASV). This paper proposes to investigate in this direction.

In this paper, four different recognizers are designed to deal with the frequency domain. They are defined as follows:

– FB: full band representing the frequency band 300–4000 Hz.

– SB1, SB2, SB3: subbands representing frequency bands 300–1600 Hz, 1100–3100 Hz, and 2500–4000 Hz, respectively. They are composed of eight coefficients each.

In contrast to the dynamic information defined in Section 2.2, these bands will be referred to as static bands and named S-FB, S-SB1, S-SB2, and S-SB3 in the next sections.

2.2. Dynamic Information

2.2.1. Fundamentals

Soong and Rosenberg have shown in [38] that short-term spectrum information (denoted as "static") and information related to the evolution of these spectra (denoted as "dynamic") are complementary and that both of them may improve speaker verification performance. This study has also revealed that dynamic information is more robust to channel variation between training and testing conditions than static information.

In the use of dynamic information, various issues have to be addressed: the length of the temporal window and the choice of appropriate front-end processing or/and modeling.

– *Length of the temporal window.* The extraction of dynamic information requires a temporal window wide enough to capture, for instance, inter-phoneme transitions or coarticulatory phenomena. On the other hand, a wide temporal window can make the characterization of the dynamic information more troublesome due to the computational complexity involved and to redundant/irrelevant data.

– *Parameterization.* During front-end processing, various kinds of features can be computed to deal with dynamic information. On one hand, they can result from an explicit extraction of dynamic features. Time-derived instantaneous features (delta and delta–delta coefficients) are the most classical ones. They represent the speed and the acceleration of spectral components [21]. Tracking the time trajectories of spectral components [29] and tracking the formants [33] are also parameterizations proposed in the literature. On the other hand, some approaches do not rely on explicit extraction of dynamic features during

[2] Experiments were carried out on TIMIT and NTIMIT databases.

the paramaterization. Concatenating successive instantaneous feature vectors is a solution investigated in [23, 25, 27]. In this case, the extended feature vectors convey information of both static and dynamic nature. Finally, classic instantaneous features can also be a potential parameterization to deal with dynamic information. In this case, the dynamic information has to be exploited by the model itself.

– *Suitable models for dynamic information.* Predictive models have been widely proposed in the literature to deal with dynamic information. They consist in modeling the dynamics of the speech signal by predicting a frame from the previous frames in the signal. Depending on the prediction function, one can find predictive models such as autoregressive vectorial models [32] (ARVM), neural networks [3], time delay neural networks (TDNN) [5]. While predictive approaches have been designed to model the "dynamic" of the speech signal, i.e., to exploit dynamic information intrinsically within the speaker models, other techniques have been used to handle dynamic information extracted during the front-end processing. For instance, time-derived features are usually fed into classical models such as HMM and its variants (mono Gaussian models (MGM), Gaussian mixture models (GMM), etc.). Another suitable approach consists in increasing the size of the feature vectors used for the model (HMM, GMM, etc.) by taking a larger temporal window into account (parameterization based on the concatenation of several successive feature vectors as seen previously) [1, 18].

2.2.2. Choice of a Suitable "Dynamic" Approach

Fredouille and Bonastre report in [18] that a temporal window extended up to 100 ms contains sufficient data to exploit dynamic information in the context of speaker recognition. Depending on this window width, some points can be discussed:

– ARVM is an attractive approach since it is able to model speech signal evolution intrinsically. As model complexity increases with the temporal window size and/or the model order, only second order ARV models are usually experimented. However, Magrin-Chagnolleau *et al.* demonstrate in [28] that a second order ARV model tends to indirectly extract speaker characteristics of a static nature rather than of a dynamic nature.

– In [3, 5], the approaches based on neural networks or TDNN classifiers tend to perform quite well in ASI. However, the adaptation of such techniques for ASV does not seem to be easily feasible. Indeed, few studies have shown satisfactory results for this kind of technique in the framework of speaker verification. This is likely due to the large amount of training data required or to the involved computation complexity.

– Time-derived features are able to handle the overall information contained in such a temporal window easily. Their performance has been well demonstrated in the literature [21]. However, this front-end processing can be seen as a process of compression of the information contained in a large temporal window. Indeed, the information related to a large temporal window is

summarized into a single delta or delta–delta feature vector. Therefore, it can be assumed that part of the useful information is lost during compression.

– Tracking the time trajectory is an interesting approach. But, similarly to the time-derived features, this technique relies on a compression process of dynamic information which might result in a loss of relevant information.

In contrast with the time-derived features or the tracking of the time trajectory, which tend to extract one kind of dynamic information, this paper proposes to retain the overall information contained in a temporal window of 100 ms. A selection procedure will be in charge of retrieving the useful dynamic information afterward. In this context, the concatenation of successive feature vectors (associated with statistical modeling) seems to be the approach most suitable to respond to this request. It is detailed in the following section.

2.2.3. "Dynamic" Approach Proposal

Let W be the size of the temporal window required to exploit dynamic information and t the size of a speech signal frame. Each extended feature vector is composed of W/t concatenated frames (i.e., W/t successive spectra).

These extended feature vectors are assumed to convey dynamic information as well as redundant and/or irrelevant information. A selection procedure is then applied to retrieve speaker-specific information. The choice of an optimal set of parameters is an issue widely discussed in the literature. In this paper, the ascendant method proposed in [12] (a variant of the well-known knock-out method [37]) has been chosen for its reduction of computational complexity. This selection approach consists in evaluating, at each iterative step, all the subsets $S(n+1)$ built by adding to $S(n)$ one coefficient among the p unused ones (at the first iteration, $S(n) = \{\}$ and p is equal to the number of potential coefficients N). Then the best subset is selected and the procedure is repeated ($n = n+1$ and $p = p-1$) until all coefficients are finally added.

This selection approach has been associated with a selection criterion optimized [3] for the tasks concerned in this paper. This procedure results in reduced feature vectors representing the subset of optimum parameters. Finally, a state-of-the-art GMM approach is applied on these reduced feature vectors to build speaker models.

In this paper, four different recognizers are designed to deal with dynamic information. They are defined as follows:

– D-FB: a dynamic full band composed of 144 coefficients. [4]

– OD-SB1, OD-SB2, OD-SB3: three optimized dynamic subbands [5] in which the selection [6] procedure has been applied. In average, the rate of selection of coefficients is 67% for females and 64% for males.

[3] The optimum subset of coefficients, estimated on a separate population, tends to maximize the ratio

$$\frac{\text{likelihood of true speaker } X}{\max_{Y_i}(\text{likelihood of speaker } Y_i \text{ different from } X)}.$$

[4] For reasons of computational complexity, no selection has been performed for this recognizer.

[5] The frequency domain is split into three subbands as detailed in Section 2.1.

[6] The selection procedure is gender-dependent.

3. BLOCK-SEGMENTAL APPROACH

The use of multiple recognizers, each designed to process one kind of speaker-specific information, seems to be a good way to deal with the intrinsic complexity of the speech signal. However, due precisely to the differences between the recognizers, implementing this architecture is not an easy task. When it comes to merging various recognizer results, difficulties arise from the heterogeneousness of these outputs, as well as from the nature of the considered speaker-specific criteria. Furthermore, the recognizers can require different amounts of data, located at different places in the speech signal, to yield meaningful scores. In this case, if the system were to output one score per recognizer and per signal frame, this would introduce substantial redundancy within the output streams of some recognizers, while making others unable to produce valid answers.

A solution consists in using a segmental approach. The speech signal is split into several temporal segments, corresponding to zones where a given recognizer can find useful information. This means avoiding output redundancy for this recognizer, results being yielded only once for each segment (of course, another benefit is the significant decrease in the number of values to be merged). It also allows robustness to be improved by deleting the zones where no pertinent information is found [8].

But such a segmentation must obviously be made on a recognizer-by-recognizer basis, as the usefulness of information varies depending on the criterion considered. Besides causing more computational complexity, this implies knowing for each recognizer how to retrieve segments of useful information. Furthermore, the introduction of asynchronicity between the outputs of the various recognizers strongly increases the complexity of the fusion process. Finally, while output normalization is still needed for this fusion, another kind of normalization is required to merge the segment scores for each recognizer.

A simpler segmentation-based method is proposed here, as a compromise between performance and complexity: the speech signal is split into temporal blocks of fixed length (hence the name "block-segmental approach"), the segmentation being the same for all the recognizers.

This arbitrary segmentation avoids having to deal with synchronization problems in the fusion of the recognizer scores, while still decreasing the total amount of values to merge and preserving the ability to delete noninformative zones when accumulating the block scores. Having the same amount of data within every block also eases normalization between blocks for each recognizer.

This segmentation scheme takes its full sense when combined with a normalization method capable of bringing all scores—for all blocks and all recognizers—into the same numerical domain. This is done by taking into account, for each block/recognizer pair, both the intrinsic performance of the recognizer and the amount of useful information within the block.

Fusion and pruning then become easier tasks, thanks to the complexity being concentrated in the normalization step.

3.1. Description of the Block-Segmental Approach

Let B_i be one of the blocks resulting from the fixed-length segmentation introduced above.

B_i is composed of T speech signal frames $(y_t)_{i\times T<tr\leq(i+1)\times T}$.

Score $S_n(B_i \mid \mathcal{X})$ of this block for recognizer n ($n \in \{1, \ldots, N\}$), given speaker model $\mathcal{X}$, is defined as

$$S_n(B_i \mid \mathcal{X}) = f(S_n(y_{i\times T+1} \mid \mathcal{X}), \ldots, S_n(y_{(i+1)\times T} \mid \mathcal{X})), \tag{1}$$

where $S_n(y_t \mid \mathcal{X})$ is the score of frame y_t yielded by recognizer n, stemming from the first step—Norm 1_n—of the normalization process (see Section 3.2).

The value (T) to be given to the block length obviously depends on the nature of function f. However, other factors have to be considered. Larger blocks mean fewer results to merge, thus making the fusion step easier. They should also lead to more reliable block scores, on which the normalization function may be more efficient. On the other hand, a small block size helps to carry out a more precise pruning of noninformative zones.

In this paper, a geometrical mean is used as function f, with a block length (T) of 30 frames:

$$f(S_n(y_{i\times T+1} \mid \mathcal{X}), \ldots, S_n(y_{(i+1)\times T} \mid \mathcal{X})) = \left(\prod_{j=1}^{T} S_n(y_{i\times T+j} \mid \mathcal{X})\right)^{1/T}. \tag{2}$$

The choice of the geometric mean is guided by the assumption that all the frames are homogeneous within a block and have been uttered by the same speaker.

Given this block score, the second step normalization function (Norm 2_n) returns probability $P_n(B_i \mid \mathcal{X})$ (for recognizer n) that the speaker corresponding to $\mathcal{X}$ has uttered speech block B_i:

$$P_n(B_i \mid \mathcal{X}) = \text{Norm}\,2_n(S_n(B_i \mid \mathcal{X})). \tag{3}$$

The normalization process is detailed in Section 3.2.

The last step is to merge the normalized scores of all blocks for all recognizers. The scores have to be fused both

- between the various recognizers (vertical axis) and
- along the temporal axis, between the blocks (horizontal axis).

The possibility of temporal accumulation of the block scores allows the interrecognizer fusion to be delayed. This fusion process can be achieved at each temporal block, as in this paper, or delayed until the last block is processed (all the intermediate fusions are conceivable). The interrecognizer fusion may be based on any classical fusion technique such as arithmetical mean, geometrical mean, NBest, NWorst, or majority vote [7, 26].

In this paper, a simple arithmetic mean is applied. A single score $P(B_i \mid \mathcal{X})$ is first computed for each block by averaging the outputs of the various

recognizers:

$$P(B_i \mid \mathcal{X}) = \frac{1}{N} \sum_{n=1}^{N} P_n(B_i \mid \mathcal{X}). \tag{4}$$

Then the resulting scores $P(B_i \mid \mathcal{X})$ are merged. While geometric mean would be a logical choice here, arithmetic mean has proved to perform better during speaker detection tests on a development dataset. This performance difference, which may be explained by the lower influence aberrant block scores have on the final score in the case of arithmetic mean, leads us to use the latter for the speaker detection task. In the case of the Two Speakers and Speaker Tracking tasks, the block scores are sorted prior to the merging, thus discarding the problem of aberrant values (see Section 6 for details) and allowing us to use geometric mean.

3.2. Normalization

In the proposed architecture, besides dealing with classical speech variability issues, the normalization method takes the behavior of the various recognizers into account. Consequently, a normalization function is defined for each recognizer.

The normalization process consists of two steps, $\mathrm{Norm}\,1_n$ and $\mathrm{Norm}\,2_n$, combining two normalization techniques used in ASV [20].

The first step consists in computing a ratio between the likelihood of hypothesis H_0: "the speech signal was uttered by the speaker"—summarized by the similarity measure between the speech signal and the speaker model—and the likelihood of hypothesis H_1: "the speech signal was uttered by another speaker." Hypothesis H_1 relies on a generic anti-speaker model and is often represented by a world model in the literature.

Thus, a world-model-based likelihood ratio is computed [11, 35] for each frame,

$$S_n(y_t \mid \mathcal{X}) = \frac{L_n(y_t \mid \mathcal{X})}{L_n(y_t \mid \bar{\mathcal{X}})}, \tag{5}$$

where $L_n(y_t \mid \mathcal{X})$ (resp. $L_n(y_t \mid \bar{\mathcal{X}})$) is the measure of similarity between signal frame y_t and speaker model $\mathcal{X}$ (resp. world model $\bar{\mathcal{X}}$) yielded by recognizer n.

The second step corresponds to the $\mathrm{Norm}\,2_n$ function introduced in Eq. (3) (Sect. 3.1). Close to a MAP (Maximum A Posteriori) normalization [31], it consists in replacing a block score with the a posteriori probability of this score being a target score (as opposed to a nontarget or impostor score). This probability is computed according to the Bayes rule, using both

– a priori probabilities for target and impostor scores; these probabilities actually describe the test conditions (and are obviously the same for all recognizers);

– a posteriori probability density functions of target and impostor scores for each recognizer; these functions are estimated on a separate development data set.

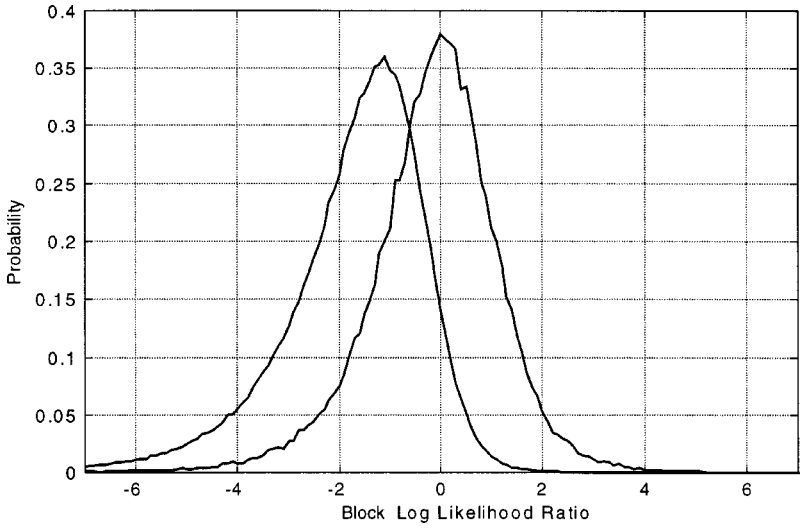

FIG. 2. World+MAP normalization. Probability density functions of target- and non-target-block log likelihood ratios obtained on the development data set.

The probability $P_n(\text{tar} \mid S_n)$ of a score S_n—yielded by recognizer n—being a target score is then defined by

$$P_n(\text{tar} \mid S_n) = \frac{P_n(S_n \mid \text{tar}) \times P(\text{tar})}{P_n(S_n \mid \text{tar}) \times P(\text{tar}) + P_n(S_n \mid \text{imp}) \times P(\text{imp})}, \tag{6}$$

where $P_n(S_n \mid \text{tar})$ and $P_n(S_n \mid \text{imp})$ are the probabilities for score S_n given the a posteriori probability density functions of target and impostor scores for recognizer n, and $P(\text{tar})$ and $P(\text{imp})$ are the a priori probabilities of target and impostor scores.

For better understanding of the World+MAP normalization, Figs. 2–4 depict the different phases of the World+MAP normalization (see [19] for more details). Indeed, Fig. 2 provides the a posteriori probability density functions (pdf) of target and impostor (nontarget) block score for a given recognizer. These pdfs have been yielded on a development data set, defined by the ELISA consortium and extracted from the NIST 98 evaluations. Target and impostor scores are world-model-based log likelihood ratios computed on blocks $(S_n(B_i \mid \mathcal{X})$, as defined in Eq. (1)).

From these pdfs, a Bayesian normalization function can be estimated as mentioned above. In this context, the a priori probabilities for target $(P(\text{tar}))$ and non target $(P(\text{non}))$ scores are set to 0.1 and 0.9, respectively, in order to match closely the NIST evaluation conditions. Figure 3 illustrates this normalization function. It is defined by three main parts:

– the first part, in which target representative probabilities, greater than $P(\text{tar})$ (a priori target score probability), are assigned to log likelihood ratios;

– the second part, in which all the probabilities are smaller than $P(\text{tar})$, referring to nontarget scores;

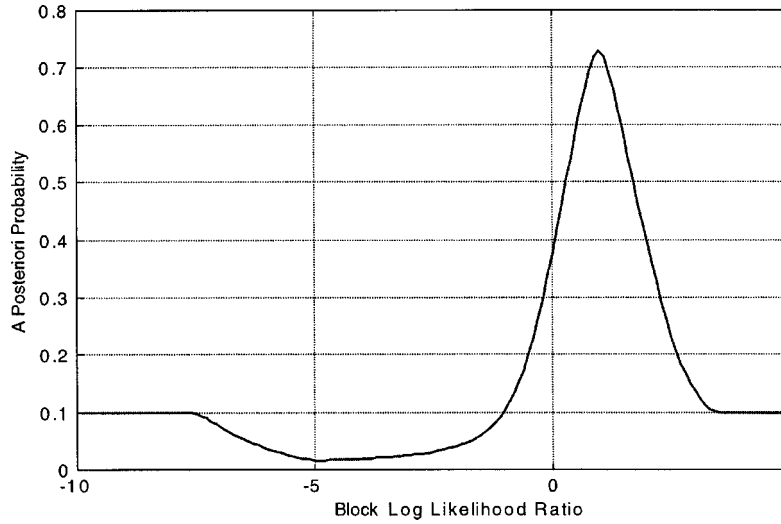

FIG. 3. World+MAP normalization. Normalization function estimated from block log likelihood ratio pdfs obtained on the development data set.

– the last part, in which probabilities are assigned to the a priori target score probability, referring to the noninformative log likelihood ratios (e.g., unusual ratio values).

The normalization process is assessed by applying the Bayesian normalization function on a separate evaluation data set. This set has a similar structure to the development data set (used to learn the normalization function) and is also defined by the ELISA consortium. From the Bayesian normalization,

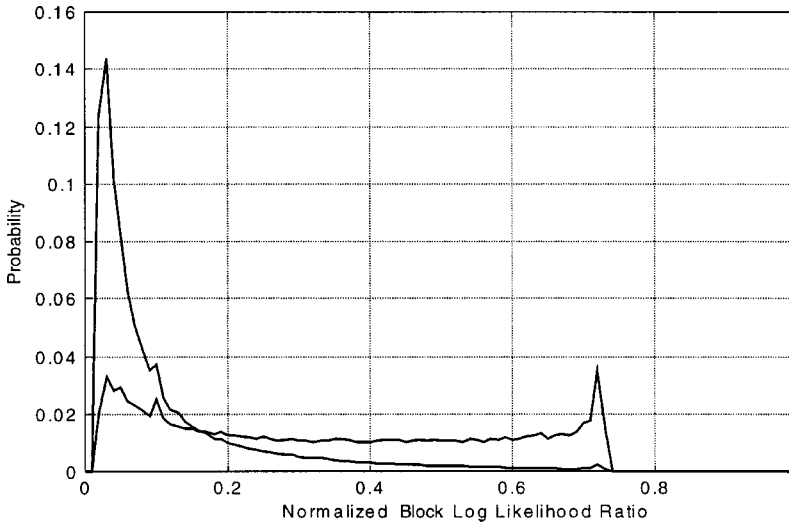

FIG. 4. World+MAP normalization. Probability density functions of normalized target and nontarget scores obtained on the development data set.

probabilities have been yielded. Figure 4 provides the resulting normalized target and nontarget score pdfs obtained on the development data set since they are quite similar to those obtained on the evaluation data set. This figure demonstrates that:

– As expected, the normalized nontarget scores are mainly concentrated in the range $[0; P(\text{tar})]$;

– The normalized target scores can be divided into two main parts. In the first one, corresponding blocks are correctly labeled as belonging to a target speaker with a confidence dependent on the probability value. The second part relates to ratios with probabilities smaller than $P(\text{tar})$ which should be labeled as nontarget scores and may correspond to error-prone blocks due to a lack of speaker-specific information.

4. EXPERIMENTAL FRAMEWORK

Experiments reported in this paper have been conducted in the context of the NIST/NSA 99 speaker recognition evaluations.

The AMIRAL system, described in the next section, has been developed by the LIA in the framework of the ELISA consortium [16], which the LIA is a member of. Like all software modules developed in the ELISA framework, it will be made available to all the consortium members for the next NIST evaluations.

4.1. AMIRAL System

AMIRAL is a system dedicated to automatic speaker recognition. It has been implemented by the LIA in order to respond to the main tasks of ASI and ASV. The potentiality of AMIRAL relies on the association of a multirecognizer architecture defined in Section 2 with the block-segmental technique described in Section 3. The next sections will present the other modules of the AMIRAL system illustrated by Fig. 1.

4.1.1. Front End Processing

AMIRAL integrates the parameterization module developed within the ELISA consortium.[7] Among the various parameterization techniques proposed by this module, AMIRAL uses the classical cepstrum analysis. The speech signal is characterized, every 10 ms, by 16 cepstrum features, derived from a filter bank analysis on a 32.5 ms-wide window. A Cepstral Mean Subtraction (CSM) is applied afterwards on each cepstrum vector in order to minimize the degradation of speech signal due to the various transmission channels.

No additional feature related to the energy of the speech signal or delta and delta–delta coefficients is used.

[7] The ELISA consortium is composed of European research laboratories working on a shared reference platform for the evaluation of speaker recognition systems. These labs are ENST (France), EPFL (Switzerland), IDIAP (Switzerland), IRISA (France), LIA (France), RIMO—Rice (USA) and Mons (Belgium), RMA (Belgium), and VUTBR (Czech Republic).

4.1.2. Speaker Modeling

AMIRAL applies different statistical voice-modeling techniques. In this paper, each speaker is characterized by a mono-state model. The speaker models implied are the so-called Gaussian mixture models (GMM) [34, 36] trained with a classical EM (Expectation–Maximization) algorithm [14] based on the maximum likelihood principle. Let y_t be a p-dimensional feature vector of speech signal uttered by speaker $\mathcal{X}_s$. The mixture density is defined as

$$p(y_t \mid X_s) = \sum_{i=1}^{M} p_s^i \mathcal{N}\left(y_t, \mu_s^i, \Sigma_s^i\right), \tag{7}$$

where p_s^i and $\mathcal{N}(y_t, \mu_s^i, \Sigma_s^i)$ are respectively the mixture weights which satisfy the constraint $\sum_{i=1}^{M} p_s^i = 1$ and the ith Gaussian density summarized by mean vector μ_s^i and covariance matrix Σ_s^i.

In this paper, each speaker-dependent Gaussian mixture model is composed of either 16 components summarized by a full covariance matrix each (static recognizers) or 128 components summarized by a diagonal covariance matrix each (dynamic recognizers). In any case, the gender- and handset-dependent world models, used for the normalization, have the same characteristics as speaker models. On the other hand, gender-dependent world models have been used during speaker model training to initialize the EM algorithm.

Finally, the similarity measure between an incoming feature vector, representing a speech frame, and a model consists in estimating the likelihood for the speech frame of being emitted by the model.

4.1.3. Normalization

It has to be noticed that no *znorm*- and *hnorm*-like normalization is used here. As opposed to this kind of normalization, the normalization used here (see Section 3.2) is independent of the test clients; i.e., it makes no use of information from client models (particularly, no per-client impostor distributions are computed).

4.1.4. Decision Thresholds

The decision strategies vary depending on the task considered. For the Speaker Detection and Two Speakers tasks, it consists in a simple comparison between the final score, issued from the block score merging (see Section 3.1) and a threshold. For the Speaker Tracking task, a more complex strategy is applied (see Section 6) that also involves comparison with a threshold.

Decision thresholds are speaker-independent, i.e., no threshold adaptation to the speakers is carried out, neither in the normalization (see Sections 4.1.3 and 3.2) nor during the decision step.

All the thresholds have been tuned a priori using data set defined by the ELISA consortium, made of one-speaker segments extracted from the NIST 98 evaluation database. Thresholds for Speaker Detection and Two Speaker tasks have been optimized for 30-s-long tests. The threshold for Speaker Tracking has been optimized for tests of 3 s (which is assumed to correspond to the average duration of a speaker utterance in a conversation).

4.2. NIST/NSA 99 Speaker Recognition Evaluations

Since 1996, the National Institute of Standards and Technology (NIST) has coordinated evaluation campaigns of text-independent speaker recognition systems over the telephone.

Each year, the evaluation conditions aim at focusing on certain specific issues. Since 1996, focus has been particularly on the effect of handset types. Since 1999, multispeaker recordings have also been of interest.

The research sites (academic or industrial laboratories) involved in these evaluation campaigns have to supply, for each speaker recognition test, a binary detection decision as well as a decision score associated with it.

4.2.1. Evaluation of the Speaker Recognition Tasks

In the first years, the NIST/NSA evaluations focused on the task of speaker verification. It consists in determining whether a test speech segment has been uttered to a specified speaker and is implied in a context of conversational telephone speech. In 1998–1999, two new tasks have been introduced. The first one, called Two Speakers, is quite close to speaker detection but deals with speech segments containing both sides of a telephone call rather than the speech of a single speaker. The second one, called Speaker Tracking, consists in retrieving speech segments of a known speaker involved in a conversation.

A detection cost function (DCF) is used to measure system performance for the different tasks. This DCF, based on the (binary) detection decision, is defined as follows:

$$C_{\text{det}} = C_{\text{fr}} \times P_{\text{fr}} \times P_{\text{Target}} + C_{\text{fa}} \times P_{\text{fa}} \times P_{\overline{\text{Target}}}. \tag{8}$$

C_{fr} (resp. C_{fa}) is the relative cost of a false rejection (resp. a false acceptance). P_{Target} (resp. $P_{\overline{\text{Target}}}$) is the a priori probability of a client trial (resp. an impostor trial).

P_{fr} and P_{fa} are the measured false rejection and false acceptance rates. For the special task of Speaker Tracking, P_{fr} and P_{fa} are estimated as follows:

$$P_{\text{fr}} = \frac{\#\text{ of target frames labeled as nontarget}}{\#\text{ of target frames}} \tag{9}$$

$$P_{\text{fa}} = \frac{\#\text{ of nontarget frames labeled as target}}{\#\text{ of nontarget frames}}. \tag{10}$$

In addition, the decision scores are used to produce a Detection Error Tradeoff (DET) curve [30], illustrating the tradeoff of false acceptances and false rejections.

4.2.2. Switchboard Database

The 99 evaluation campaign was carried out on a subset of the Switchboard II corpus composed of 230 male and 309 female speakers. For each speaker, about 2 min of speech were available as training data. Stemming from two different sessions, these data were used to estimate a model per speaker. Similarly, various test segments whose length ranged from 2 s to 1 min were extracted for each speaker.

TABLE 1

Test Conditions: Number of Tests Depending on the Task

Task	Number of tests
Speaker detection	37,620
Two speakers	37,906
Speaker tracking	4000

On the other hand, a subset of the 98 evaluation campaign corpus was extracted to estimate gender- and handset-dependent world models. This subset was composed of speech signal recordings 30 s long uttered by 100 female and 100 male speakers.

Table 1 reports the number of tests depending on the task.

5. COMPARISON OF VARIOUS MULTIRECOGNIZER ARCHITECTURES

A series of experiments has been conducted to evaluate the multirecognizer approach in the framework of One Speaker detection. First, individual recognizers have been tested to estimate their own performance. Then, different multirecognizer architectures have been experimented with. These experiments aim at demonstrating the following aspects:

- the pertinence of the information coupled with each recognizer;
- the potentiality of the proposed multirecognizer-based approach;
- the potentiality of the dynamic approach proposed in this paper.

5.1. Performance of Individual Recognizers

Figures 5 and 6 give the DET curves of static (S-FB, S-SB1, S-SB2, and S-SB3) and dynamic (D-FB, OD-SB1, OD-SB2, OD-SB3) recognizers, respectively. Full band recognizers outperform subband recognizers (S-FB compared to S-SB1, S-SB2, S-SB3, and D-FB compared to OD-SB1, OD-SB2, and OD-SB3). Static and dynamic full band recognizers without selection obtain similar performance (a slight improvement of performance is observed for the dynamic full band). Finally, S-SB1, S-SB3, OD-SB1, and OD-SB3 have worse performance than S-SB2 and OD-SB2.

These results show that both static and dynamic subbands yield poor performance. This behavior tends to demonstrate that the use of frequency subbands is not pertinent in this context of conversational speech in a real telephony environment.

5.2. Comparison of Different Architectures

Multiple architectures have been designed from different combinations of recognizers and tested to compare their performance. They are defined as follows:

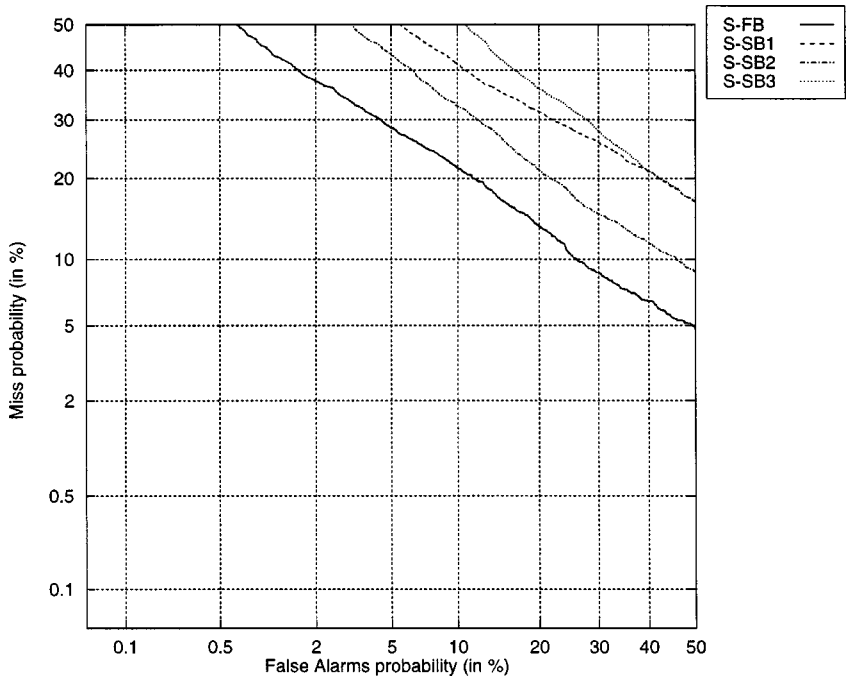

FIG. 5. Static recognizers. Comparison of the performance obtained by individual static recognizers for "one speaker" detection.

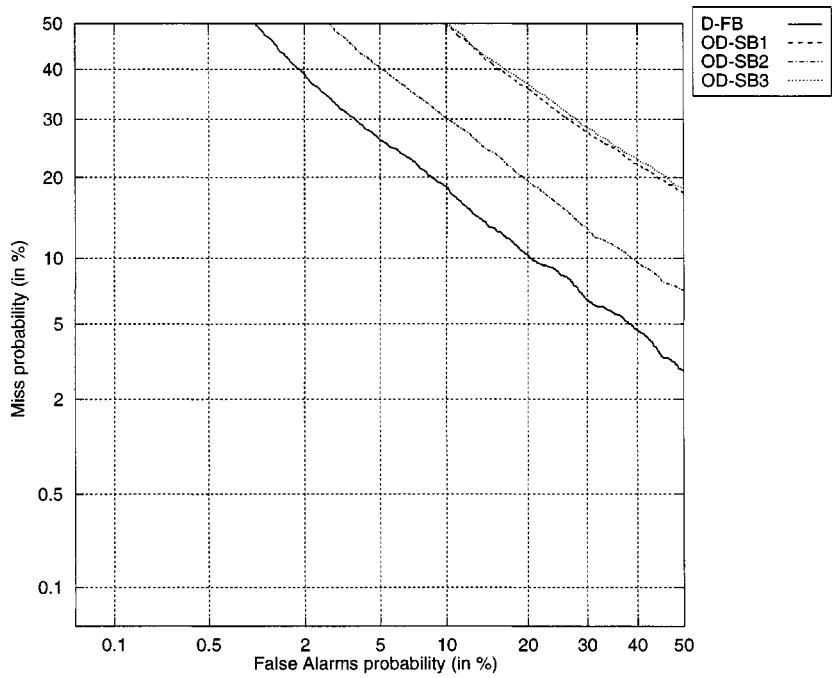

FIG. 6. Dynamic recognizers. Comparison of the performance obtained by individual dynamic recognizers for "one speaker" detection.

(i) Static architecture "SFB": referring to the single static full band (frequency recognizer only). This is the reference architecture.

(ii) Hybrid architecture "SFB+DSB": composed of a static full band (frequency recognizer only) and three dynamic subbands (combination of frequency and dynamic recognizers).

(iii) Dynamic architecture "DFB+DSB": composed of a dynamic full band and three dynamic subbands (combination of frequency and dynamic recognizers).

The choice of these architectures has been guided by various aspects. First of all, the static architecture SFB has been chosen for its low cost in parameters and resources. Indeed, this architecture is simply a monorecognizer 16 component GMM based on static cepstrum features only. The hybrid architecture SFB+DSB has been presented to evaluate the correlation between static and dynamic information. In addition, this architecture has allowed the use of dynamic subbands to be quantified in terms of performance. Finally, the dynamic architecture "DFB+DSB" has been introduced in order to observe the correlation between dynamic full band and subbands.

Figure 7 provides the DET curves obtained by each of these architectures. The dynamic architecture (DFB+DSB) gives the best performance. The hybrid

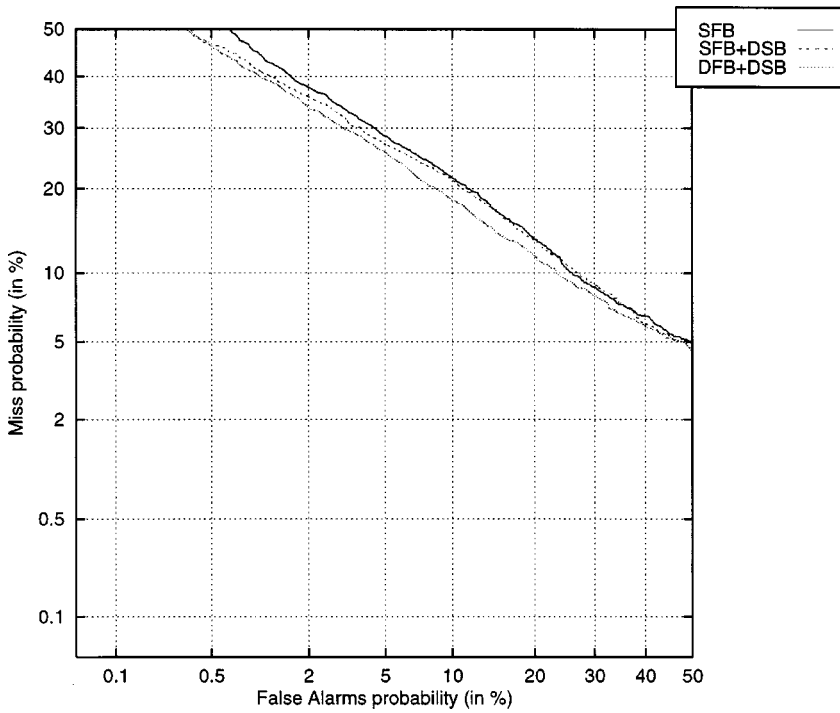

FIG. 7. Static versus dynamic. Comparison of different multirecognizer architectures—SFB (static architecture), SFB+DSB (hybrid architecture), DFB+DSB (dynamic architecture)—for "one speaker" detection.

architecture, formed of static and dynamic recognizers, slightly outperforms the static architecture.

Comparing the DET curve of the dynamic architecture with the curve of the dynamic full band (DFB) recognizer only (provided in Fig. 6) shows comparable performance. DFB slightly outperforms DFB+DSB. This tends to question the use of dynamic subbands or more specifically either the frequency subbands implied in this architecture or the quality of the selection approach, applied on the dynamic subbands uniquely.

5.3. Dynamic Approach versus Time-Derived Parameters

Time-derived parameters are classically used to extract dynamic information from speech signals. Therefore, this approach has been tested in the same conditions as those used for the dynamic full band recognizer D-FB (without selection of optimum coefficients) to compare their performance.

5.3.1. Time-Derived Coefficients

To test the approach based on time-derived coefficients, called later the delta/delta–delta approach, static vectors composed initially of 16 cepstrum features have been extended with 16 first-derived cepstrum features (delta coefficients) and 16 second-derived cepstrum features (delta–delta coefficients)

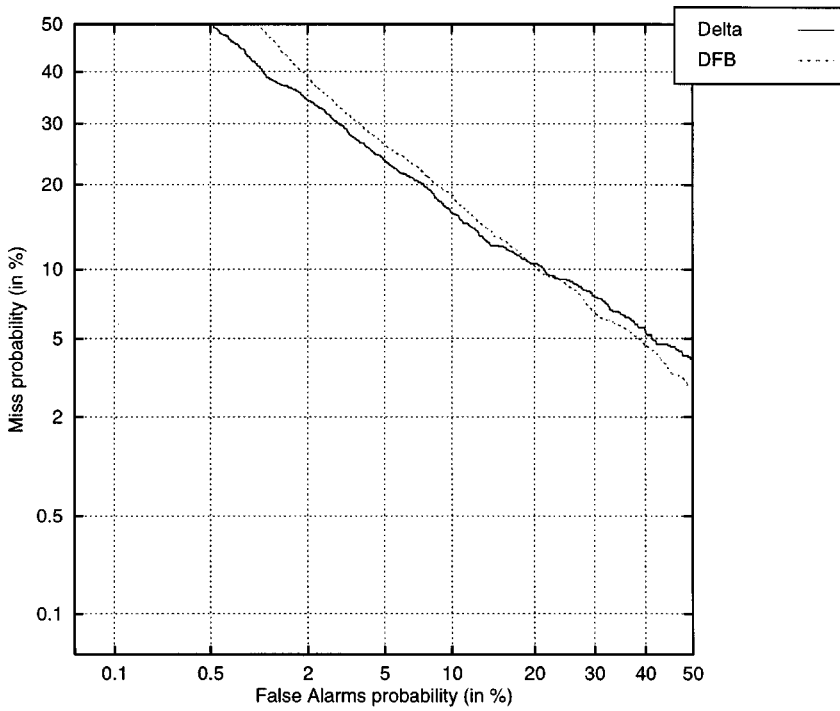

FIG. 8. Dynamic versus delta/delta–delta. Comparison of two dynamic approaches, one based on the concatenation of successive feature vectors and the other on time-derived coefficients for "one speaker" detection.

during parameterization. As for static vectors, GMM speaker models have been estimated.

5.3.2. Results

Figure 8 provides the DET curves of both delta/delta–delta and DFB based approaches. Few differences in performance between the two approaches have been observed. Delta/delta–delta slightly outperforms DFB, whereas the latter uses all the coefficents within the dynamic window. Therefore, there appears to be redundancy between these coefficients. Carrying out a reduction of the number of coefficients should lead to a good compromise.

6. THE USE OF AMIRAL FOR THE TWO SPEAKERS AND SPEAKER TRACKING TASKS

As mentioned in Section 4.2, two new tasks have been recently introduced into the NIST/NSA evaluation campaigns, dealing with speech segments involving multiple speakers:

– The Two Speakers task is an extension of speaker detection to multi-speaker speech segments containing both sides of a telephone call.

– The Speaker Tracking task is a bit more complex as, besides detecting the presence of a given speaker within a conversation, the goal here is to determine precisely the boundaries of his utterances.

The AMIRAL system has been only adapted to achieve the Speaker Tracking task since Two Speakers is considered as the first step of Speaker Tracking.

The flexibility brought by the block-segmental approach has made this adaptation relatively easy. In fact, the whole process is the same as for speaker detection, until the end of the interrecognizer fusion. The difference lies in the fusion of the block scores and the decision step.

The fusion/decision method used here, called Sorted Weighted Geometric Mean (SWGM), allows simultaneous determination of whether and when the considered speaker's voice is present in the conversation. It consists of four steps:

– The blocks are first sorted according to their scores.

– The second step is the search for the optimal block subset on which the speaker detection is to be carried out. This is the one whose weighted geometric mean is maximal. The purpose of the weighting function is to emphasize the number of blocks in the considered subset, thus compensating the score value decrease due to the geometric mean. In this paper, the weighting function has been determined empirically, by observing the behavior of the SWGM on about 30 tests. SWGM is defined as $m \times a^{b/n}$, where m is the geometric mean and n is the cardinal of the score set. The estimated parameters are $a = 0.1$ and $b = 1$.

– The value of the weighted mean for the selected block subset is then compared to a predetermined threshold. Given whether it is greater or lower, the target speaker is said to be present or not in the document. In the case of the Two Speakers task, this decision represents the final result.

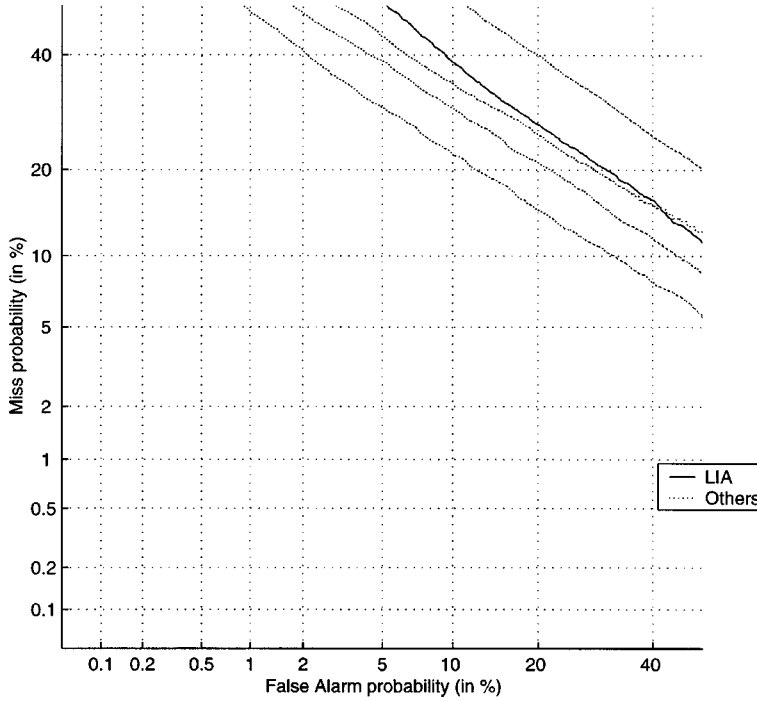

FIG. 9. Two speakers. DET curves for all the participants in the NIST 99 evaluations.

– In the case of Speaker Tracking, the last step consists (if the target speaker presence has been detected) in extending the selected set by adding blocks (best scores first) until the weighted mean gets under the decision threshold. This extended set of blocks is finally attributed to the target speaker.

Figure 9 depicts the DET curve of the AMIRAL system (monorecognizer, SFB-based system) and of all the other NIST participants for the Two Speakers task. While correct in view of the novelty of the task, these results stand definitely behind those of the best systems.

Although the SWGM method has still to be improved (with the use of optimization techniques to refine the wieghting function), poor performance mainly originates from the underlying architecture (static full band monorecognizer with no hnorm-like normalization). Indeed, results for this architecture, using respectively arithmetic mean and SWGM for merging, applied to the One Speaker speaker verification task (Fig. 10), show little performance difference between both merging functions.

Results for the Two Speakers tasks should therefore improve with the use of the best AMIRAL architecture (i.e., using multiple recognizers and dynamic information) as the base system for SWGM.

Figure 11 shows the DET curve of the AMIRAL system (monorecognizer, SFB-based system) for the Speaker Tracking task along with that of a more classical speaker tracking system.

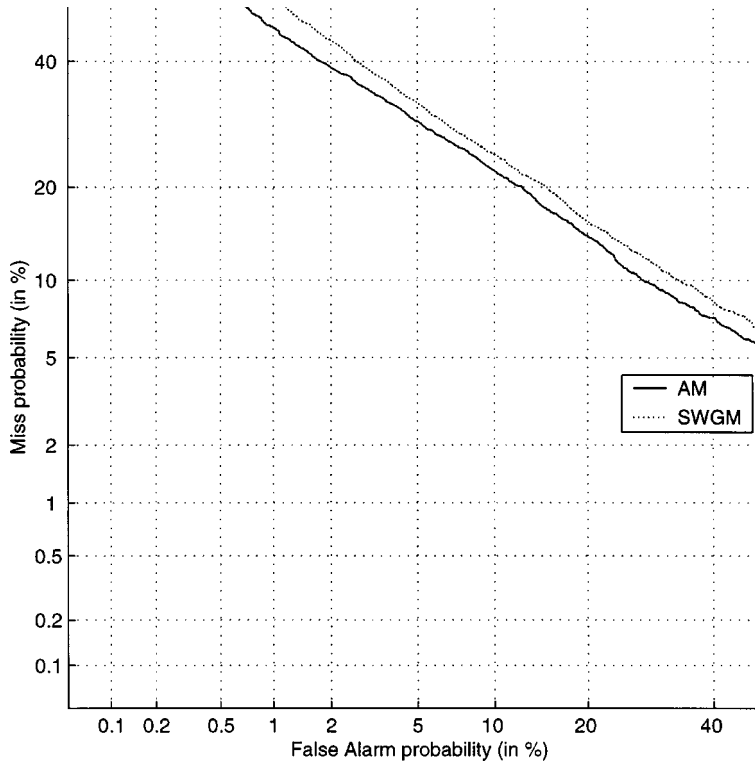

FIG. 10. One speaker detection. Comparison between arithmetic mean and SWGM for block score merging.

This second system is based on a speaker change detection method [13] which splits the signal into single-speaker speech segments without using any speaker model. Speaker verification is then carried out on the resulting speech segments using the AMIRAL system to determine which segments belong to the target speaker.

Compared to this classical method, the original approach proposed here presents the advantage of merging the segmentation and decision steps. Indeed, rather than carrying out a blind segmentation and using information available about the target speaker only for the decision phase, it is preferable to exploit it during the segmentation process as well, thus making this task considerably easier.

Furthermore, this method leads to a unique decision, taken globally on all the segments where the target speaker is likely to appear. The verification process being carried out on a larger amount of data should give decisions more accurate than ones taken on a segment-by-segment basis.

However, one drawback of the current implementation consists in the precision of the segmentation process: indeed, in order to follow the underlying block segmentation, the AMIRAL system only yields segments with length a multiple of 0.3 s. This has to be compared to the precision of 1/100 s of the speaker change detection system.

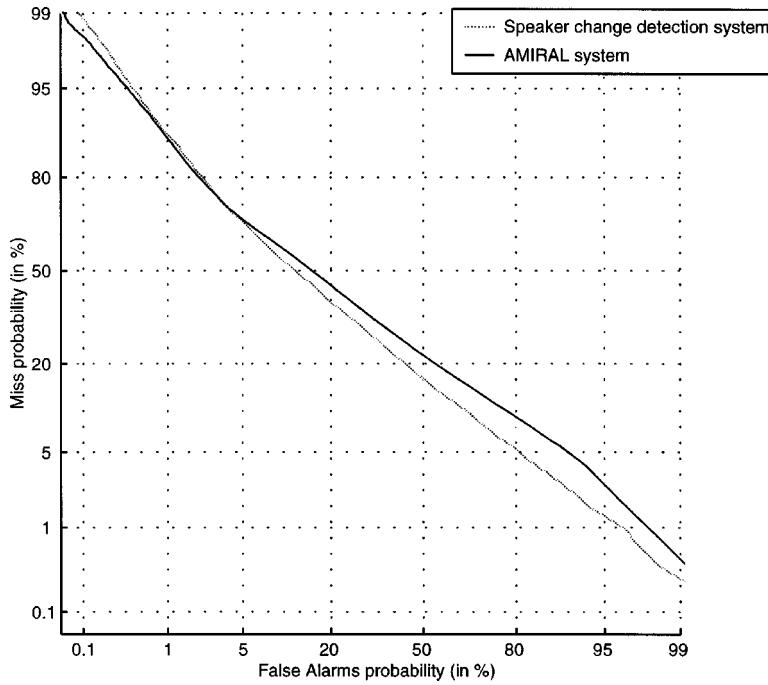

FIG. 11. Speaker Tracking. Comparison of two systems for the Speaker Tracking task. The first approach is based on the AMIRAL system and the second on a classical speaker change detection technique.

Nevertheless, despite this precision difference, both systems appear to perform roughly similarly (the AMIRAL system has a slightly higher EER, but performs better according to the NIST detection cost function). Given the simplicity of the implementation realized here, this has to be considered encouraging for this method, as several improvements should come from both the SWGM method and the AMIRAL recognizer architecture.

7. DISCUSSION AND PERSPECTIVES

Considering the complexity of the speech signal, multirecognizer architectures tend to be a fairly attractive solution. By devoting each classifier to a particular piece of information, this kind of approach is able to improve speech/speaker recognition systems. Classically, handling multiple classifiers is a complex task. It also addresses one of the most discussed issues in the literature: the fusion issue. This additional complexity makes the multirecognizer system less flexible for adaptation to new recognition tasks (Speaker Tracking).

The AMIRAL system proposed in this paper is based on a block-segmental multirecognizer architecture. In contrast with classical multirecognizer approaches, it remains quite simple. Indeed, thanks to the block-segmental approach, this system permits, the underlying complexity to be shifted toward the block-normalization process. A block-normalization function is computed inde-

pendently for each recognizer and takes its intrinsic performance into account. This normalization scheme produces homogeneous Bayesian scores which make the fusion of multiple recognizers simpler. This proposed solution has demonstrated its potentiality. However, this technique is mainly constrained by the availability of additional training data which have to be representative of both the task and conditions under consideration. It can be noticed that a similar constraint (related to the conditions) applies to classical systems for the likelihood domain normalization [22].

Finally, the flexibility induced by the block-segmental approach allows an easy and fast adaptation of the AMIRAL system to the task of Speaker Tracking. Results provided in Section 6 show that the AMIRAL system is able to perform as well as a classical speaker change detection approach.

During the NIST/NSA 1999 campaign, several versions of the AMIRAL system were evaluated. A basic monorecognizer block-segmental version based on a cepstral parameterization without dynamic features (no delta or delta–delta parameters), a state-of-the-art 16-Gaussian-based GMM summarized by full matrices and no threshold normalization [22](HNorm, ZNorm,...) obtained quite satisfactory performance. This observation shows the interest of the proposed block-segmental approach and of the World+MAP normalization. It also demonstrates how challenging it is to improve the performance of a full band monorecognizer system. The complete AMIRAL system (DFB+DSB) shows a slight increase of performance if compared to this basic version. This small improvement of performance is rather disturbing regarding the potentialities of the proposed approaches. These results bring into question the choice of subbands in these particular evaluation conditions.

Consequently, studying new parameterizations to provide several full band recognizer seems to be a more convenient approach.

Finally, threshold normalizations such as ZNorm and HNorm have proved their efficiency for the ASV task [22]. It is therefore interesting to introduce a similar principle within the block-normalization approach.

7.1. Frequency Bands

Results presented in Section 5.1 show that performance differs significantly from one subband to another. The overall performance of the subbands is rather weak if compared with the full band performance (averaged EER of $\approx$28% for the subbands against $\approx$15% for the full band). It can be noticed that subband SB2, which represents a frequency band from 1100 to 3100 Hz, performs best. This result does not match the observations made in [8] regarding the localization of the useful part of information for the task of ASI on TIMIT and NTIMIT databases.

The difference between the tasks (ASV vs ASI) as well as the difference between the databases (read speech vs real conversational speech) considered in the two studies in question could explain this result.

7.2. Dynamic Information

The results obtained in Section 5.3 highlight an important aspect: using the concatenation of successive feature vectors (illustrated by DFB) to handle

dynamic information can lead to performance similar to delta and delta–delta features.

In this experimental context, no selection process is coupled with the dynamic approach proposed. Therefore, further investigations have to be conducted to estimate the potential gain in performance induced by a selection of the useful part of information.

On the other hand, concatening several feature vectors leads to having both the static and the dynamic information present in the temporal window. It should be interesting to split these two classes of information to decrease the redundancy in the case of a multirecognizer architecture. It could be assumed that the selection procedure discussed previously will perform this separation intrinsically. However, an explicit means of splitting these two classes should reinforce the robustness of the system.

7.3. Static and Dynamic Architectures

The results obtained in Section 5.2 emphasize the well-known potentialities of dynamic information:

– Complementarity of static and dynamic information. The hybrid architecture obtains better performance than the static architecture. Dynamic recognizers supply additional speaker-specific information since performance is improved. This particularity has been demonstrated in [38] with delta and delta–delta coefficients but negated in [6].

– Robustness of dynamic information. The dynamic architecture presents better performance than the hybrid architecture. This result is very close to [21, 38] which demonstrate that time-derived coefficients are more resistant to linear channel mismatch between training and testing than instantaneous features.

REFERENCES

1. Afify, M. and Haton, J.-P., Non-parametric segment models for automatic speaker identification. In *Proc. on Speaker Recognition and its Commercial and Forensic Applications RLA2C'98, Avignon,* April 1998, pp. 68–71.
2. Allen, J. B., How do humans process and recognize speech? *IEEE Trans. Acoust. Speech Signal Process.* **2** No. 4 (1994), 567–577.
3. Artières, T. and Gallinari, P., Neural models for extracting speaker characteristics in speech modelization systems. In *Eurospeech'93,* 1993, pp. 253–257.
4. Auckenthaler, R. and Mason, J. S., Score normalisation in a multi-band speaker verification system. In *Proc. on Speaker Recognition and its Commercial and Forensic Applications RLA2C'98, Avignon,* 1998, pp. 102–105.
5. Bennani, Y. and Gallinari, P., On the use of TDNN extracted features information for talker identification. In *ICASSP'91,* 1991.
6. Bernasconi, C., On instantaneous and transitional spectral information for text-dependent speaker verification, *Speech Commun.* **9** (1990), 129–139.
7. Besacier, L., *Un modèle parallèle pour la reconnaissance Automatique du Locuteur.* Ph.D thesis, Université d'Avignon et des Pays de Vaucluse, April 1998.
8. Besacier, L. and Bonastre, J.-F., Frame pruning for speaker recognition. In *ICASSP'98, Seattle,* May 1998.
9. Besacier, L., Bonastre, J.-F., and Fredouille, C., Localization and selection of speaker specific information with statistical modeling, *Speech Commun.* (1999), to be published.

10. Bourlard, H. and Dupont, S., A new ASR approach based on independent processing and recombination of partial frequency bands. In *ICSLP'96, Philadelphia,* October 1996.
11. Carey, M. J. and Parris, E. S., Speaker verification using connected words, *Inst. Acoust.* **14** (1992), 95–100.
12. Charlet, D. and Jouvet, D., Optimizing feature set for speaker verification. In *AVBPA'97,* Lecture Notes in Computer Science, Vol. 1207, Springer-Verlag, Berlin/New York, 1997, pp. 203–210.
13. Delacourt, P., Kryze, D., and Wellekens, C. J., Detection of speaker changes in an audio document. In *Eurospeech'99, Budapest,* Vol. 3, 1999, pp. 1195–1198.
14. Dempster, D., Larid, N., and Rubin, D., Maximum likelihood from incomplete data via EM algorithm, *J. Roy. Statist. Soc.* **39** (1977), 1–38.
15. Duchnowsky, P., *A New Structure for Automatic Speech Recognition.* Ph.D. Thesis, Massachussets Institute of Technology, 1993.
16. ELISA consortium, The ELISA'99 Speaker Recognition and Tracking Systems. In *Workshop on Automatic Identification Advanced Technologies,* October 1999.
17. Fletcher, H., *Speech and Hearing in Communication.* Krieger, New York, 1953.
18. Fredouille, C. and Bonastre, J.-F., Use of dynamic information with second order statistical methods in speaker identification. In *Proc. on Speaker Recognition and Its Commercial and Forensic Applications RLA2C'98, Avignon,* April 1998, pp. 50–54.
19. Fredouille, C., Bonastre, J.-F., and Merlin, T., Segmental normalization for robust speaker verification. In *Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere,* May 1999.
20. Fredouille, C., Bonastre, J.-F., and Merlin, T., Similarity normalization method based on world model and a posteriori probability for speaker verification, In *Eurospeech'99, Budapest,* Vol. 2, pp. 983–986, September 1999.
21. Furui, S., Cepstral analysis for automatic speaker verification, *IEEE Trans. Acoust. Speech and Process.* **2** No. 2 (1981), 254–272.
22. Gravier, G. and Chollet, G., Comparison of normalization techniques for speaker verification. In *Proc. on Speaker Recognition and its Commercial and Forensic Applications RLA2C'98, Avignon,* April 1998, pp. 97–100.
23. Hattori, H., Text-independent speaker recognition using neural networks. In *ICASSP'92,* 1992, pp. 153–156.
24. Hermansky, H., Tibrewala, S., and Pavel, S., Towards ASR on partially corrupted speech. In *ICSLP'96, Philadelphia,* October 1996.
25. Higgins, A. L. and Wohlford, R. E., A new method of text-independent speaker recognition. In *ICASSP'86, Tokyo,* 1986.
26. Kittler, J., Li, Y. P., Matas, J., and Ramos Sanchez, M. U., Combining evidence in multimodal personal identity recognition systems. In *Proc. AVBPA,* Lecture Notes in Computer Science, (Bigün *et al.*, Eds.), Springer-Verlag, Berlin/New York, 1997, pp. 327–334.
27. Konig, Y., Heck, L., Weintraub, M., and Sonmez, K., Nonlinear discriminant feature extraction for robust text-independent speaker recognition. In *Proc. on Speaker Recognition and Its Commercial and Forensic Applications RLA2C'98, Avignon,* April 1998, pp. 72–75.
28. Magrin-Chagnolleau, I., Wilke, J., and Bimbot, F., Further investigation on AR-vector models for text-independent speaker identification. In *ICASSP'96, Atlanta,* May 1996.
29. Magrin-Chagnolleau, I. and Durou, G., Time-frequency principal components of speech: application to speaker identification. In *Eurospeech'99, Budapest,* pp. 759–762, September 1999.
30. Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M., The DET curve in assessment of detection task performance. In *Eurospeech'97,* Vol. 4, pp. 1895–1898, 1997.
31. Matsui, T. and Furui, S., Likelihood normalization for speaker verification using a phoneme- and speaker-independent model, *Speech Commun.* (1995), 109–116.
32. Montacié, C. and Le Floch, J.-L., Discriminant AR-vector models for free-text speaker verification. In *Eurospeech'93, Berlin,* 1993.
33. Paoloni, A., Pierucci, P., and Ragazzini, S., Improving automatic formant tracking for speaker identification. In *Proc. on Speaker Recognition and Its Commercial and Forensic Applications RLA2C'98, Avignon,* April 1998, pp. 24–27.
34. Reynolds, D. A., Identification and verification using Gaussian Mixture speaker models, *Speech Commun.* (1995), 98–108.
35. Reynolds, D. A., Comparison of background normalization methods for text-independent speaker verification. In *Eurospeech'97,* 1997, pp. 963–966.

36. Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., Speaker verification using adapted Gaussian mixture models, *Digital Signal Process.* **10** (2000), 19–41.
37. Sambur, M. R., Selection of acoustic features for speaker identification, *IEEE Trans. Acoust. Speech Signal Process.* **23** No. 2 (1975), 176–178.
38. Soong, F. K. and Rosenberg, A. E., On the use of instantaneous and transitional spectral information in speaker recognition, *IEEE Trans. Acoust. Signal Speech Process.* **36** No. 6 (1988), 871–879.