Subband Architecture for Automatic Speaker Recognition

Laurent BESACIER, Jean - François BONASTRE

Abstract

We present an original approach for automatic speaker identification especially applicable to environments which cause partial corruption of the frequency spectrum of the signal. The general principle is to split the whole frequency domain into several subbands on which statistical recognizers are independently applied and then recombined to yield a global score and a global recognition decision. The choice of the subband architecture and the recombination strategies are particularly discussed. This technique had been shown to be robust for speech recognition when a narrow band noise degradation occur. We first objectively verify this robustness for the speaker identification task. We also study which information is really used to recognize speakers. For this, speaker identification experiments on independent subbands are conducted for 630 speakers, on TIMIT and NTIMIT databases. The results show that the speaker specific information is not equally distributed among subbands. In particular, the low-frequency subbands (under 600Hz) and the high-frequency subbands (over 3000Hz) are more speaker specific than middle-frequency ones. In addition, experiments on different subband system architectures show that the correlations between frequency channels are of prime importance for speaker recognition. Some of these correlations are lost when the frequency domain is divided into subbands. Consequently we propose a particularly redundant parallel architecture for which most of the correlations are kept. The performances obtained with this new system, using linear recombination strategies, are equivalent to those of a conventional fullband recognizer on clean and telephone speech. Experiments on speech corrupted by unpredictable noise show a better adaptability of this approach in noisy environments, compared to a conventional device, especially when pruning of some recognizers is performed.

Résumé

Nous proposons une nouvelle approche pour la reconnaissance automatique du locuteur, particulièrement bien adaptée à des environnements causant une dégradation partielle du signal de parole. L'originalité de l'approche vient du découpage de l'espace fréquentiel en sous-bandes sur lesquelles des reconnaisseurs statistiques sont indépendamment appliqués puis recombinés pour obtenir un score de reconnaissance global et prendre une décision finale. Nous discutons plus particulièrement le choix de l'architecture en sous-bandes et les stratégies de recombinaison. Ce type d'approche s'est déjà montré robuste en reconnaissance de la parole lorsque le signal est corrompu par un bruit bande étroite. Cette robustesse est vérifiée objectivement pour la tâche d'identification du locuteur. Nous étudions également quelle information est réellement utilisée pour reconnaître les locuteurs. Pour cela, des tests d'identification sont réalisés sur des sous-bandes indépendantes pour 630 locuteurs des bases TIMIT et NTIMIT. Les résultats montrent que l'information utile à la caractérisation du locuteur est surtout présente en basse fréquence (f<600 Hz) et dans les hautes fréquences (f>3000 Hz) alors que les performances d'identification baissent nettement pour les bandes intermédiaires. Des expériences sur différentes architectures en sous-bandes montrent également que les corrélations entre canaux fréquentiels sont importantes pour la tâche de reconnaissance du locuteur. Certaines de ces corrélations sont perdues lorsque le domaine fréquentiel est divisé en sous-bandes. En conséquence, nous proposons une architecture parallèle particulièrement redondante, pour laquelle la plupart des corrélations sont conservées. Les performances obtenues avec ce nouveau système sont équivalentes a celle obtenues par un reconnaisseur conventionnel « bande totale » sur des signaux de parole propre (normal et téléphone). Des expériences sur de la parole bruitée montrent une meilleure capacité adaptative de cette approche, surtout lorsque la stratégie de recombinaison autorise l'élagage de certains reconnaisseurs.

Subband Architecture for Automatic Speaker Recognition

Laurent BESACIER[1], Jean - François BONASTRE[1]

(1) LIA/CERI - Agroparc - 339, chemin des Meinajaries BP 1228 - 84911 Avignon Cedex 9 (France)

laurent.besacier@lia.univ-avignon.fr , jean-francois.bonastre@lia.univ-avignon.fr

List of symbols

$\bar{x}$	mean vector of speaker X
X	covariance matrix of speaker X
$\{y_t\}_{1\le t\le N}$	sequence of N vectors uttered by speaker Y
$G(y_t / X)$	likelihood of acoustic vector y_t
$G^k(y_t / X)$	likelihood of acoustic vector y_t on the k-th subband
$\overline{L_X}(y_1^N)$	average log-likelihood of utterance $\{y_t\}_{1\le t\le N}$
$\overline{L_X}^k(y_1^N)$	average log-likelihood of utterance $\{y_t\}_{1\le t\le N}$ on the k-th subband
$\mu(X, y_t)$	minus-log-likelihood of y_t
$\mu^k(X, y_t)$	minus-log-likelihood of y_t on the k-th subband
$\mu(X, y_i^j)$	average minus-log-likelihood of utterance $\{y_t\}_{i\le t\le j}$; equivalent to a similarity measure between speaker X and speaker Y on the speech segment [i,j]
$\mu^k(X, y_i^j)$	average minus-log-likelihood of utterance $\{y_t\}_{i\le t\le j}$ on the k-th subband
$\rho^k(X, y_i^j)$	position where the reference model of speaker X appears in the ordered list of neighbours of test utterance $\{y_t\}_{i\le t\le j}$ on the k-th subband
$\delta^k(X, y_i^j)(m)$	abstract score : $\delta^k(m)=0$ if model of speaker X is one of the *m* nearest neighbours of test utterance $\{y_t\}_{i\le t\le j}$ on the k-th subband ; $\delta^k(m)=1$ else
$s(X, y_i^j)$	score after recombination of scores $s^k(X, y_i^j)$ *with* $s = \{\mu, \rho, \delta\}$
s_1	equal weighting recombination
s_2	non discriminant weighting
s_3	discriminant weighting
s_4	'all or nothing' weighting

Pages #21

Tables #5

Figures #5

Keywords speaker identification, subband architecture, recombination, parallel model, correlations, noisy speech.

1. Introduction

Most of the degradations affecting automatic speaker recognition systems are related to the variability of signal characteristics from trial to trial [9]. The speech variations result from speaker himself (intra-speaker variability), from differences in recording or transmission conditions, and from surrounding noise.
Lately, many papers have been dedicated to the acoustic mismatch due to different training and testing environments in telephone speech (different handsets and channels)[12][20][22]. The recent arrival of mobile radiotelephony raises new challenges : low bit rate speech coding, noise cancellation and adaptation (since the noise characteristics may change over time) [11]. For the moment, the lack of GSM recordings has limited the studies to the effects of speech coding algorithms on speaker verification [15]. As far as the problems due to noise are concerned, their simulation can be considered. Although this is not representative of what happens in real applications, it is worth starting to work with this artificial material until real condition databases are available.

The methods used to cope with speech variation problems can be classified into three levels : signal level, feature level and model level [10]. The approach proposed here for speaker recognition concerns the latter. The general principle is to split the whole frequency band into several subbands on which statistical recognizers are independently applied and then recombined to yield global scores and a global recognition decision. In fact, when a speech signal is partly degraded by a frequency selective noise, a part of the speech spectrum may still carry valid information. However, a typical signal representation used in automatic speaker recognition consists of a series of feature vectors, each vector representing the entire short-term frequency spectrum at a given time instant. Even only one or a few corrupted elements in the feature vector can lead to severe degradation of the recognition performance. The independent processing of the subbands should allow to reject the degraded ones and to take the final decision with a subset of the best subband scores.

Multiband models for speech recognition have been proposed in [5], [6] and [21]. Beyond the task which is not the same, our approach differs in two main points :
- First, in order to reasonably compare the efficiency of our subband approach with a full band one, we wanted to strictly handle the same number of acoustic parameters for both approaches. Actually, in some studies related to multiband speech recognition, the number of LPC parameters at the input of the systems, is bigger for the multiband recognizer than for the conventional full band recognizer. In that case, one may not be able to conclude that the potential improvement is due to the multiband paradigm or simply due to the increase in the number of coefficients. For that reason, we decided to use a set of filter bank coefficients which is the same for both multiband and conventional systems. Then, the input vector is either considered globally or split into subvectors each corresponding to a frequency subband.
- The second difference was influenced by an observation made in [21] where it is reported that the improvement in performance obtained with a non linear recombination using a Multi Layer Perceptron (MLP) « should not be attributed only to the multiband paradigm ». In fact, it is said that when the MLP is trained on independent data to re-classify the outputs of the baseline (fullband) system, the improvement is similar to that achieved by MLP merging of

the subbands. Consequently, in order to objectively compare conventional and multiband approaches, we decided to use only linear recombination rather than a MLP.

In spite of these differences, some advantages of multiband speech recognition may be generalized to the speaker recognition task :
- the system may be robust in the case of speech corrupted by a noise affecting a limited number of subbands,
- different recognition techniques might be applied to different subbands,
- the subband approach is well suited for taking advantage of parallel architectures,
- the most speaker specific subbands can be localized with this approach.

Finally, the efficiency of the subband method significantly depends on two factors :
(1) *the architecture of the subband-based system :* selection of the most critical subbands for the recognition task ; optimal division of the whole frequency domain (number of subbands, size of subbands),
(2) *the recombination of the output of each subband recognizer :* recombination level, recombination strategies, fusion of multiple decisions.
Some considerations concerning these two aspects are respectively presented in *section 2* and *section 3*. Experimental conditions are described in *section 4* ; the systems proposed in this paper are tested for 3 different speech qualities : normal speech, telephone speech and speech corrupted by simulated noise (narrow band and full band noise). In *section 5*, experiments with different subband architectures are reported in order to understand the phenomena that happen in speaker recognition and to know which information is really used to recognize speakers (Most discriminant subbands ? Importance of the correlations between subbands ?). The robustness of the multiband approach against narrowband noise is also verified in this section. The conclusions drawn in *section 5* lead us to propose a particularly redundant architecture which is experimented in *section 6*. Finally, further advantages of the multiband approach for speaker recognition are discussed in *section 7*.

2. Subband Architecture

2.1 Basic Principle

The subband-based speaker recognition system can be seen as a combination of multiple recognizers (one for each subband) associated to a decision module which performs the recombination of each subband recognizer output. *Figure 1* illustrates the general principle of our multiband system.

2.2 Mono Gaussian Model

In this section, the mono-gaussian modeling, which is the starting point of our subband model, is presented. This gaussian modeling is more precisely described in [4].
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 :

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \quad \text{et} \quad X = \frac{1}{M}\sum_{t=1}^{M}(x_t - \bar{x})(x_t - \bar{x})^T \qquad \textit{(Equation 1)}$$

Similarly, for a speech signal uttered by speaker Y, a sequence of N vectors $\{y_t\}_{1\le t\le N}$ 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 :

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

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

$$\overline{L_X}(y_1^N) = \frac{1}{N}\log G(y_1 \ldots y_n | X) = \frac{1}{N}\sum_{t=1}^{N}\log G(y_t | X) \qquad \textit{(Equation 3)}$$

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

$$\underset{X}{Arg\,max}\, G(y_t/X) = \underset{X}{Arg\,min}\, \mu(X, y_t) \qquad \textit{(Equation 4)}$$

we have

$$\mu(X,y_t) = -\log G(y_t / X) \qquad \textit{(Equation 5)}$$

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

$$\mu(X,Y) = \mu(X, y_1^N) = \frac{1}{N}\sum_{t=1}^{N}\mu(X, y_t) = -\overline{L_X}(y_1^N) \qquad \textit{(Equation 6)}$$

This measure is equivalent to the standard gaussian likelihood measure (asymmetric μ_G) defined in [4]. This measure will be used for the experiments described in *section 6* and *5.3*. In the experiments of *section 5.1* and *5.2*, the following symmetric version of this measure (β symmetrisation [4]) is used :

$$\mu_{G_{[\beta_{MN}]}}(X,Y) = \frac{M.\mu(X,Y) + N\mu(Y,X)}{M+N} \qquad \textit{(Equation 7)}$$

2.3 Multiband Formalism

The speaker modelisation with a covariance matrix and a mean vector is particularly well adapted to the design of a multiband approach. The following 'K-subbands' model of speaker X can be obtained from the initial full-band model :

$$\mathcal{M}_X(K) = \left\{\left(X^1, \bar{x}^1\right), \ldots, \left(X^k, \bar{x}^k\right), \ldots, \left(X^K, \bar{x}^K\right), M\right\} \qquad \textit{(Equation 8)}$$

where speaker X is modeled on the k-th subband with covariance matrix X^k and mean vector $\bar{x}^k$. X^k is a sub-block of the covariance matrix X and $\bar{x}^k$ is a sub-vector of the mean vector $\bar{x}$ (X and $\bar{x}$ being computed on the whole spectral domain). Therefore, the quantities defined in *Equations 2,3,5* and *6* can be respectively written for the k-th subband :

- $G^k(y_t / X)$ likelihood of acoustic vector y_t on the k-th subband,
- $\overline{L_X}^k(y_1^N)$ average log-likelihood of utterance $\{y_t\}_{1\le t\le N}$ on the k-th subband,
- $\mu^k(X, y_t)$ minus-log-likelihood of y_t on the k-th subband
- $\mu^k(X, Y)$ similarity measure between speaker X and speaker Y on the k-th subband

From now on, we will refer to architecture Kxd when the multiband model is composed of K subbands of *d* channels each (*d*=dimension of the subband vectors). In that case, the number of channels per subband is constant and overlap between subbands is possible.

3. Recombination Strategies

3.1 Recombination Level

The recombination task can occur at different levels [7]:
- the measures obtained on each subband are combined and a final decision is taken with the results of the measure fusion,
- a partial decision is made for each subband and the final decision results from a recombination of these partial decisions.

More precisely, the information level at the output of each recognizer can be of three types [17]. *Figure 2* illustrates the relative scores assigned to speakers for each information level :

(1) 'distance' level : a minus-log-likelihood $\mu^k(X, y_1^N)$ is computed at the output of each k-th subband recognizer,

(2) 'sorting' level : an identification rank $\rho^k(X, y_1^N)$ is calculated at the output of each k-th subband recognizer. $\rho^k(X, y_1^N)$ is the position where the reference model of speaker X appears in the ordered list of neighbours of test utterance $\{y_t\}_{1\le t\le N}$ on the k-th subband,

(3) 'abstract' level : an 'all or nothing' score $\delta^k(X, y_1^N)(m)$ (0 or 1) is given at the output of each k-th subband recognizer, i.e. a subset of the *m* most probable speakers is selected.

Levels (2) and (3) are particularly useful when one of the recognizers is corrupted because the prejudice caused on the final score, by an extremely abnormal measure, is diminished when this measure is replaced by a rank ρ or a vote δ.

3.2 Recombination Window

Another aspect of the subband model concerns the choice of a recombination window to compute the final similarity measure between the reference model of speaker X and the sequence of N vectors $\{y_t\}_{1\le t\le N}$ of speaker Y.

Actually, the merging task can occur for different time window sizes : recombination after each frame, recombination after *T* frames, or only one recombination at the end of the whole test utterance. Whereas the window length can be variable, it will be constant in the experiments and only extreme sizes of windows will be considered :

- recombination of the measures after each frame (T=1) ; in that case, we enforce synchronization points at each frame,
- recombination procedure at the end of the test utterance (T=N) ; in that case, the subbands are processed independently till the end of the signal.

The log-likelihood of vector $(y_i)_{t+1\le i\le t+T}$ on the window (or segment) of T frames [t+1,t+T], for the k-th subband is :

$$\mu^k(\mathsf{X}, y_{t+1}^{t+T}) = -\overline{L_{\mathsf{X}}}^k(y_{t+1}^{t+T}) = -\frac{1}{T}\log G^k(y_{t+1}\ldots y_{t+T}|\mathsf{X}) = -\frac{1}{T}\sum_{i=1}^{T}\log G^k(y_{t+i}|\mathsf{X}) \qquad \textit{(Equation 9)}$$

After the recombination phase, the score for segment [t+1,t+T] is then :

$$s(\mathsf{X}, y_{t+1}^{t+T}) = F\left(s^1(\mathsf{X}, y_{t+1}^{t+T}),\ldots,s^k(\mathsf{X}, y_{t+1}^{t+T}),\ldots,s^K(\mathsf{X}, y_{t+1}^{t+T})\right) \qquad \textit{(Equation 10)}$$

where *F()* is the recombination function and s={μ, ρ, δ} (scores defined in section 3.1).

Finally, the global score of the whole test utterance of *N* frames results from the accumulation of all the segments scores :

$$s(\mathsf{X}, y_1^N) = \frac{1}{n}\sum_{t=0}^{n-1} s(\mathsf{X}, y_{tT+1}^{tT+T}) \qquad \textit{(Equation 11)}$$

where *T* is the number of frames in a segment and *n* is the number of segments in the whole test utterance so that the total number of frames in the whole test utterance is N=*n*T.

3.3 Fusion Operator

For the reasons explained in the *Introduction*, a linear merging technique, which corresponds to a weighted sum of each recognizer output, is used. Four types of weighting have been examined to recombine the scores $s^k(\mathsf{X}, y_i^j)$ with s={μ, ρ, δ} (scores defined in section 3.1) :

i) equal weighting (the same confidence associated to each subband recognizer) :

$$s_1(\mathsf{X}, y_i^j) = \frac{1}{K}\sum_{k=1}^{K} s^k(\mathsf{X}, y_i^j) \qquad \textit{(Equation 12)}$$

ii) non discriminant weighting, according to the accuracy of individual subbands (weights derived from the identification performances of the individual subbands on a tuning data set) :

$$s_2(\mathsf{X}, y_i^j) = \frac{1}{K}\sum_{k=1}^{K} w_k^{no-disc}.s^k(\mathsf{X}, y_i^j) \qquad \textit{(Equation 13)}$$

iii) discriminant weighting, a genetic algorithm is used to learn the weights by minimization of a cost function (MCE criterion) on a tuning data set :

$$s_3(\mathsf{X}, y_i^j) = \frac{1}{K}\sum_{k=1}^{K} w_k^{disc}.s^k(\mathsf{X}, y_i^j) \qquad \textit{(Equation 14)}$$

iii) 'all or nothing' weighting, we investigate here merging using a hard threshold approach by removing the least reliable subband scores ; i.e. some weights fixed to 0, some others to 1 (pruning of the recognizers which lead to abnormally high distance measures) :

$$s_4(\mathsf{X}, y_i^j) = arg\,min_1^p \left[\frac{1}{p} \sum_{k \in \{1..K\}} s^k(\mathsf{X}, y_i^j) \right] \qquad \textit{(Equation 15)}$$

4. Experimental Conditions

4.1 Databases

For the experiments, 3 different speech qualities are proposed : *normal* speech, *telephone* speech and *noisy* speech.

4.1.1 Normal Speech - TIMIT

TIMIT database [8] contains 630 speakers (438 male and 192 female), each of them having uttered 10 sentences. The speech signal is recorded through a high quality microphone, in a very quiet environment, with a 0-8 kHz bandwidth. All recordings took place in a single session (contemporaneous speech).

4.1.2 Telephone Speech - NTIMIT

The NTIMIT database [14] was obtained by playing TIMIT speech signal through an artificial mouth installed in front of the microphone of a fixed handset frame and transmitting this input signal through a telephone line. For each speaker, there are 6 different telephone lines (local or long distance network), but half of the speaker files are transmitted through the same line. The signal is sampled at 16 kHz, but its useful bandwidth is limited to telephone bandwidth (approximately 300-3400 Hz).

4.1.3 Noisy Speech

4.1.3.1 Narrow band noise - 'TIMIT_NBN'

In that case, bandpass limited noise with a central frequency of 987 Hz and a bandwidth of 365 Hz is artificially added to the original test set (TIMIT) at different SNRs. This modified version of TIMIT will be called TIMIT_NBN (for TIMIT Narrow Band Noise). Note that the SNR is estimated with the signal energy and the noise energy both calculated in the limited noise bandwidth. This type of frequency selective degradation is common to a few real noises of the NOISEX-92 database like volvo, babble, factory and destroyer-engines noises [21]. It will allow us to generalize, to the speaker recognition task, the robustness to narrowband noise already demonstrated in [5] and [21] for multiband speech recognition.

4.1.3.2 Full band noise - 'TIMIT_FBN'

In that case, a noise spread on the whole spectrum is added to the speech signal (TIMIT database). For each frame, *C* (*C*=2 or 3) frequency channels among p (dimension of the full-band acoustic vector) are randomly selected and degraded for different SNRs. This modified version of TIMIT will be called TIMIT_FBN (TIMIT Full Band Noise) since the average

noise energy of a test utterance is equally distributed on the whole spectral domain ; moreover, this type of noise is unpredictable.

4.2 Signal Analysis

The speech analysis module extracts filterbank coefficients in the following way : a Winograd Fourier Transform is computed on Hamming windowed signal frames of 31.5 ms (i.e. 504 samples) at a frame rate of 10 ms (160 samples). For each frame, spectral vectors of 24 Mel-Scale Triangular-Filter Bank coefficients (24 channels) are then calculated from the Fourier Transform power spectrum, and expressed in logarithmic scale. The central frequencies of the 24 filters are (in Hz) : 47, 147, 257, 378, 510, 655, 813, 987, 1178, 1386, 1615, 1866, 2141, 2442, 2772, 3133, 3529, 3964, 4440, 4961, 5533, 6159, 6845, 7597. Covariance matrices and mean vectors are computed from these spectral vectors. For NTIMIT, the first 2 channels and the last 7 ones are discarded since the useful bandwidth is 330-3400Hz for these data. These analysis conditions are identical to those used in [1] and [4] which will allow us to compare our results to those obtained using similar acoustic analysis and similar databases.

4.3 Training and Test Protocols

In this protocol, training or test durations are rigorously the same for each speaker. For the training of a given speaker, all 5 'sx' sentences of TIMIT (or NTIMIT) are concatenated together and the first M samples corresponding to the training duration required (6s here) are kept. Consequently, a single reference pattern is computed from exactly the same number of samples for each speaker. The silences at the beginning and the end of sentences are not removed.
For the test of a given speaker, all 'sa' and 'si' sentences (5 in total) are randomly concatenated together and blocks of N samples corresponding to the test duration required are extracted until there is not enough speech data available (limited to a maximum number of blocks per speaker). So the test patterns are computed from exactly the same number of samples for each speaker. All the tests are made within the framework of text-independent closed-set speaker identification.
From now on, if we refer to experiments on ‘normal’ speech, training and test materials are both extracted from TIMIT ; if we refer to ‘telephone’ speech, training and test materials are both extracted from NTIMIT ; finally if we refer to ‘noisy’ speech, training material is extracted from TIMIT and tests materials from TIMIT_NBN or TIMIT_FBN.
Table 1 compares the performances (obtained with the same speaker identification measure) of the conventional « full sentences » protocol (results taken from [4]) and of our « same durations » protocol. It shows that the order of magnitude of performances is approximately the same for both protocols.

5. Architecture and speaker specific information

5.1 Experiments on Isolated Subbands

Experiments : speaker identification tests are independently conducted on 21 subbands consisting of four consecutive channels with band-overlap (subband 1 : channels 1 to 4 ,

subband 2 : channels 2 to 5... , ... subband 21 : channels 21 to 24). The similarity measure used is the measure defined in *Equation 7* and applied to each subband.
Figure 3 shows the results obtained on each isolated subband for 6s training/3s test on TIMIT and NTIMIT (630 speakers).

Results :
- Large differences between subbands are observed (5% to 25% recognition rates on TIMIT).
- Experiments on TIMIT show that the low-frequency subbands (f<600Hz) and the high-frequency subbands (f>3000Hz) are more speaker specific than middle-frequency ones.
- This confirms the drastic performance decrease generally observed on NTIMIT for which the most critical subbands are removed (channels 1-2 and 18-19-20-21-22-23-24) because of the bandlimiting (300-3400 Hz).
- The identification rates are also lower on NTIMIT for the subbands between 300Hz and 3400 Hz. This could be due to the telephone network noise and to signal distortions.

5.2 Different subband architectures

Experiments : the choice of an optimal division of the frequency domain seems to be crucial for any subband approach. In this section, experiments on different subband system architectures are presented. The independent processing and the recombination of partial frequency bands have been conducted for different divisions of the frequency domain : 12 subbands of 2 channels ; 8 subbands of 3 ; 6 subbands of 4 ; 4 subbands of 6 ; 3 subbands of 8 and 2 subbands of 12. In this experiment, the channels in a subband are either consecutive or crisscrossed and there is no overlap between subbands. For instance, the first 6-dimensional subband is made up of channels 1,2,3,4,5,6 (consecutive channels) or artificially made up of channels 1,5,9,13,17,21 (crisscrossed channels), as shown in *Table 2*. In fact, the information at the input of the system is the same for consecutive or crisscrossed channels, but it's organized differently per recognizers.
For this experiment, the final recombination score corresponds to the mean of the distance measures computed on each subband (μ_1). There is only one recombination at the end of the whole test utterance (T=300). *Table 2* shows the speaker identification results obtained independently and after this basic recombination for a 4x6 architecture (consecutive and crisscrossed channels). Table *3* shows speaker identification results obtained after recombination for different divisions of the frequency domain. We show also the number of parameters used to model a speaker in each particular architecture (1 mean vector of size d and 1 covariance matrix of size d^2 for each subband).

Results :
- In *Table 2*, the independent results obtained with 6-dimensional subbands consisting of consecutive channels confirm the lack of speaker specific information observed for the middle frequencies since the identification results are only 8.5% for channels 7 to 12 against 30.9% for channels 1 to 6 and 33.1% for channels 13 to 18.
- The identification rates independently obtained on subbands having 6 crisscrossed channels are approximately equal (*Table 2*). In this case, the information conveyed by different subbands is regularly distributed because each band is composed of channels each representing distant parts of the whole spectral domain.

- The recombination results obtained with consecutive channels are far better than those obtained with crisscrossed channels (*Table 3* : 73.1% against 29% for 3-dimensional subbands). These results show that the correlations between close channels are important for the speaker recognition task when second-order statistical methods are used. Actually, the correlations between a channel and its close neighbors are discarded when a subband is composed of crisscrossed channels.
- The differences of recombination performances between consecutive and crisscrossed channels logically decrease when the subband size is bigger since there are less close correlations lost through crisscrossing in this case.
- We also note that the recombination performances decrease when the subbands become smaller and more numerous (whether the channels are crisscrossed or consecutive). In fact, when the subbands are smaller, less parameters are used to model a speaker (*Table 3).*
- In the case of consecutive channels, the deleted parameters correspond to the correlations between distant channels. However, good results are still obtained with widely reduced sets of parameters in this case. The increase in the number of subbands, combined with the reduction of their size, can be seen as an increasing approximation on the covariance matrices ; the extreme case being 24 subbands of 1 channel each, which is equivalent to a diagonal covariance matrix.

Summary :

-1- the loss of correlations induced by the multiband architecture lead to a speaker identification performances reduction when a basic recombination is performed,

-2- however, the lost correlations between distant channels have been proved to be less important for speaker identification than the correlations between close channels,

-3- thus, subbands of consecutive channels allow a reasonable architecture since the lost correlations are the less important ones.

5.3 Robustness against narrow band noise

Robustness against narrow band noise of multiband speech recognition has already been demonstrated in [5] and [21]. In this section, this property is verified for the speaker identification task.

Experiments : a 6x4 architecture is used ; thus, a speaker model is made up of 6 subbands consisting of 4 consecutive channels each : subband 1 : channels 1 to 4 , subband 2 : channels 5 to 8... , ... subband 6 : channels 21 to 24. The recombination window is the complete test utterance (i.e. only one recombination after 3s, T=300). *Figure 4* shows the speaker identification performances of different merging techniques on speech corrupted with bandpass limited noise at different SNRs.

Best refers to the best sub-band combination picked-up by 'cheating' when looking at performances on the test data. In the *p-min* technique, we select only *p* (*p*=5 here) subbands on which the distance measures between two speakers are the smallest (see μ_4 in *section 3.3*) ; therefore, abnormally high measures due to noise are eliminated. *Borda score* [19] is an alternative voting system based on identification ranks (equivalent to ρ_1). μ_1 *normalized* is a version of μ_1 where the distances computed on each subband are normalized with the sum of the distances calculated for all reference speakers on the subband concerned. So, the similarity

measures which are globally high on a noisy subband are reduced with this technique. Finally *Conventional* refers to the standard full-band Gaussian likelihood measure.
In the special case of *Majority votes* ($\delta_1(m)$) the frequency domain is divided into 21 subbands of 4 channels with band overlap (subband 1 : channels 1 to 4 , subband 2 : channels 2 to 5... , ... subband 21 : channels 21 to 24). Each subband recognizer selects the m (m=20) most probable speakers and at the end of the test utterance, the recognized speaker is the speaker which is the most frequently selected. In the case of equal scores between speakers, the test is considered as failed.

Results : the conventional full-band recognizer seriously deteriorates with the increasing noise level. The subband recognizers which use all 6 subbands (*equal weighting, accuracy weighting, genetic weighting*) deteriorate more slowly but still rather fast. The normalized version is more robust but gives poor results on clean speech. This is quite the same for the scoring methods related to voting systems (*Borda score, Majority votes*) which are robust to noise degradation but give poor results on clean data. Logically, when allowing some of the subbands to be left out (*Best*, *p-min*), the error rate for the noisy speech is greatly improved. The system is thus robust against narrow band noise. Consequently, another possibility would be to obtain some prior knowledge from the SNR estimates in the individual subbands [13] and to reject the subbands with SNRs below a certain threshold.
Note that *p-min* operator performs better when the signal is degraded which can be surprising in appearance. However, there is an explanation for this : if narrow band noise is added to the signal, the rejected subband is systematically the subband including the noise (middle frequency band in our case). Now, it has been shown in section 5.1 that middle frequency subbands are not very speaker specific. So it's better to compute the final score by always discarding a middle frequency subband (case of noisy speech) than by discarding a subband which is not the same as the tests go by (clean speech). On this subject, we demonstrated in [1] that a slight error rate reduction was obtained on TIMIT by using only 18 channels (instead of 24) corresponding to 80% of the whole frequency domain ; here again, the eliminated channels were the middle frequency ones.

6. Proposed architecture

It has been shown, in *Section 5*, that even if the multiband architecture is robust to narrowband noise, it leads to lower performances on clean speech, compared to a conventional full band system. This is mainly due to the loss of the correlations between distant channels. Consequently, we now propose an architecture for which most of these correlations are kept. A 20-dimensional subband is proposed at the input of each recognizer, which is equivalent to the fullband frequency domain from which a 4-dimensional subband (different for each recognizer) is systematically discarded.
Although this very large subband architecture may appear as very redundant, we believe that it is worth experiment it, since it was precisely highlighted by Lippmann [16] that the redundancy between subbands may be a source of human robustness to speech degradations.

6.1 Potentiality

Experiments : in order to show the potential of this new architecture, independently of the recombination strategy, the 'best' decision is built for each test using only the answers of the

subband recognizers. Actually, a test is considered as potentially successful if the target speaker is recognized on at least 1 of the *K* subbands. Of course, this scoring is indulgent ; it is however non optimal since a test is failed if a target speaker is the second most probable speaker on all subbands, whereas one can reasonably suppose that any recombination method would lead to a successful test, in this case.
The potential recognition results obtained with this scoring, for different architectures, are presented in *Table 4* for TIMIT and NTIMIT. Architectures 1x24 (TIMIT) and 1x15 (NTIMIT) refer to the conventional full band method and architectures 24x20 (TIMIT) and 15x11 (NTIMIT) refer to the proposed multi-recognizers architecture.

Results : on both databases (TIMIT and NTIMIT), the potential of the subband approach is indisputable. Of course, these identification results are not these of the system, but they have been computed only to demonstrate that the sum of information among the very large bands is greater than the information of the full band system and that the problem of recombining the recognizers is worth working on it. Now, the main issue is to find the useful information given by the sub-recognizers and to eliminate their 'bad' answers.

6.2 Robustness to unpredictable full band noise

Experiments : the recombination window size is T=1 (i.e. one recombination after each frame) and the parallel model is made up of 24 subbands consisting of 20 channels each with band overlap (architecture 24x20).
The tests are made on a 63-speaker subset of TIMIT_FBN (20 women and 43 men). The recombination techniques are the same as in *section 5.3*. The global score at the end of the whole test utterance is an arithmetic mean of the measures computed on each frame (see *Equation 11*, accumulation of frame scores). *Figure 5* shows the identification results for different subband recombination strategies.

Results : most of the cases, the subband techniques significantly outperform the conventional fullband recognizer. The best results are obtained with *p-min* (*p*=1), i.e. when the smallest similarity measure computed at the output of 1 recognizer among 24 is kept. Unlike the architecture proposed in *Section 5*, this new architecture gives results comparable (and even slightly better) to a fullband approach on clean speech (TIMIT), in addition to be robust against noise.

6.3 Telephone speech

Experiments : in this experiment, the recombination window size is T=1 (i.e. one recombination after each frame). The parallel model is made up of 15 recognizers consisting of 11 channels each (architecture 15x11). The global measure at the end of the whole test utterance is an arithmetic mean of the recombination scores computed on each frame. The speaker identification results obtained on NTIMIT (63 speakers) are presented in *Table 5*. *Conventional* refers to the standard full-band Gaussian likelihood measure computed on channels 3 to 17 (useful bandwidth).

Results : the best results are obtained with *p-min* (p=10), i.e. when the smallest similarity measures computed at the output of 10 recognizers among 24 are kept ; but this improvement is not significative. When a basic recombination strategy is used (*equal weighting*), the identification results are similar to the results of the full-band gaussian measure, whereas the other strategies give poor results.

7. Conclusion

Summary : in this paper, we presented a new speaker recognition approach based on independent processing and recombination of partial frequency subbands. This technique had already been shown to be robust for speech recognition when a narrow band noise degradation occur. We have objectively verified this robustness for the task of speaker identification. For the first time, speaker recognition experiments on independent subbands have been conducted for 630 speakers on TIMIT and NTIMIT databases. The results have shown that the speaker specific information is not equally distributed among subbands. In particular, the low-frequency subbands (under 600Hz) and the high-frequency subbands (over 2000Hz) are more speaker specific than middle-frequency ones. Experiments on different subband system architectures have proved that the correlations between frequency channels are of prime importance. Some of these correlations are lost when the frequency domain is divided into subbands ; thus a particularly redundant parallel architecture, for which most of the correlations are kept, was proposed. The performances obtained with this new system using linear recombination strategies are equivalent to those of a conventional fullband recognizer on clean and telephone speech. However, we have shown that the sum of information among the bands is greater than the information of the full band system, in spite of the important redundancy between subbands. Experiments on speech corrupted by an unpredictable noise have shown a better adaptability of this approach in noisy environments, compared to a conventional device, especially when pruning of some recognizers is performed. On this subject, further considerations on pruning can be found in [2] and [3].

Outlook : the subband method can be advantageously used to exploit dynamic characteristics of the speech signal. In fact, 'dynamic' methods, like Auto Regressive (AR) Vector models or predictive neural networks, are subject to a problem of calculation complexity when an important time window (i.e. an important order) is used. For example, in the case of ARV models, if the acoustic vectors are p-dimensional and if a q-th order ARV model is used, speakers are modeled with a Toeplitz matrix of dimension $(pq)^2$. Consequently ARV models are often limited to order 2 or 3, and their use is sometimes a disappointing experience [18]. However, if the dynamic information is treated on independent subbands, the calculation complexity is reduced, which allow to use sufficient time windows. We have recently experimented the usefulness of the subband method to extract dynamic speaker characteristics on a 100 ms time window (order q=10). The performances obtained were very promising since the dynamic method outperformed the static one for all the individual subbands.
Beyond the dynamic aspect, subband speaker models using speaker-dependent recombination strategies are an interesting issue if we assume that some subbands should perform better for certain classes of speakers than for others. These speaker models could be applied to speaker verification. In this case, the verification task could be performed on the optimal bands of the applicant speaker.

The subband approach can also be developed at the signal analysis level and different signal-processing tasks might be applied to different subbands. For instance, the length of the analysis window could be large for the low frequencies and narrow for the high frequencies (different resolutions in order to track different speech events). In this case, the speech-processing would be quasi equivalent to a wavelet transform for which analyzing objects have different properties according to their location in the time-frequency domain.
The subband approach could also be combined to a phoneme-based analytic approach if we assume that speaker-specific subbands are different from one phoneme to another. Therefore, recombination strategies would depend on the phoneme considered.
Finally, the recombination process should include knowledge about the behavior of each recognizer in the system. This behavior could be modeled, for instance, by confusion matrices.

References

[1] L. Besacier and J.F. Bonastre, "Subband approach for automatic speaker recognition : optimal division of the frequency domain", *In Proc. Audio and Video based Biometric Person Authentication,* Springer LNCS, Bigün, et. al., Eds., 1997. pp 195-202.

[2] L. Besacier and J.F. Bonastre, "Frame Pruning for Speaker Recognition". *In Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP),* Seattle, USA, May 1998.

[3] L. Besacier and J.F. Bonastre, "Time and frequency pruning for speaker identification". *In Proc. on Speaker Recognition and its Commercial and Forensic Applications (RLA2C),* Avignon, France, April 1998.

[4] F. Bimbot, I. Magrin-Chagnolleau and L. Mathan, "Second-order statistical methods for text-independent speaker identification", *Speech Communication,* n°17(1-2), pp 177-192, August 1995.

[5] H. Bourlard and S. Dupont, "Subband-based speech recognition", *In Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing,* pp 1251-1254, Munich, Germany, April 1997.

[6] C. Cerisara and J.P. Haton, "Multiband continuous speech recognition", *NATO-ASI - Computational Models on Speech Pattern Processing,* Jersey, UK, 1997.

[7] B.V. Dasarathy, *Decision fusion.* IEEE Computer Society Press 1994. Los Alamitos, California.

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

[9] S. Furui, "Recent advances in speaker recognition", *In Proc. AVBPA, Springer LNCS*, Bigün, et al., Eds., 1997. pp 237-252.

[10] S. Furui, "Recent advances in robust speech recognition", *In ESCA-NATO Workshop on Robust speech recognition for unknown communication channels*. Pont-à-Mousson, France, 17-18 Avril 1997. pp 11-20.

[11] A. Gilloire, P. Scalart, C. Lamblin and S. Proust, "Innovative speech processing for mobile terminals : an annotated bibliography", *In COST 254 Workshop on Emerging Techniques for Communication Terminals,* Toulouse (France), 7-9 July 1997.

[12] L.P. Heck and M. Weintraub, "Handset dependent background models for robust text-independent speaker recognition", *In Proc. ICASSP 97,* Munich, Germany, pp 1071-1074.

[13] H.G. Hirsch, *Estimation of noise spectrum and its applications to SNR estimation and speech enhancement* TECHNICAL REPORT TR-93-012, International Computer Science Institute, Berkeley, CA, 1993.

[14] C. Jankowski, S. Kalyanswamy, S. Basson and J. Spitz, "NTIMIT : a phonetically balanced, continuous speech, telephone bandwidth speech database", *Proc. Internat. Conf. Acoust. Speech Signal Process. '90,* New Mexico, USA, April 1990.

[15] M. Kuitert and L. Boves, "Speaker verification with GSM coded telephone speech", *In Proc. Eurospeech 97,* Rhodes, Greece, September 1997.

[16] R.P. Lippmann, "Speech recognition by machines and humans", *Speech Communication,* 22(1), pp 1-15, 1997.

[17] P. Loonis, M. Menard and J.P. Bonnefoy, "Fusion d'Informations multi-sources : Etude comparative entre une approche connexionniste dirigée et la règle orthogonale de Dempster-Shafer", *Proc. 10th RFIA*, pp 606-614, Rennes (France), 16-18 Janvier 1996.

[18] I. Magrin-Chagnolleau, J. Wilke and F. Bimbot, "A further investigation on AR-vector models for text-independent speaker identification", *In Proc. ICASSP 96*, vol. 1, pp 401-404. Atlanta, USA, 1996.

[19] H. Moulin, *Axioms of Cooperative Decision Making.* Cambridge University Press, Cambridge-New-York-Port Chester-Melbourne-Sydney 1988.

[20] D.A. Reynolds, "Speaker Identification and verification using gaussian mixture models". *In Workshop on Automatic Speaker Recognition and Verification,* pp 27-30, April 1994. Martigny (Switzerland).

[21] S. Tibrewala and H. Hermansky, "Subband-based recognition of noisy speech", *In Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing,* pp 1255-1258, Munich, Germany, April 1997.

[22] S. Van-Vuuren, "Comparison of Text independent speaker recognition methods on telephone speech with acoustic mismatch", *In Proc. ICSLP 96*, Philadelphia, USA, pp 1788-1791.

List of Figures

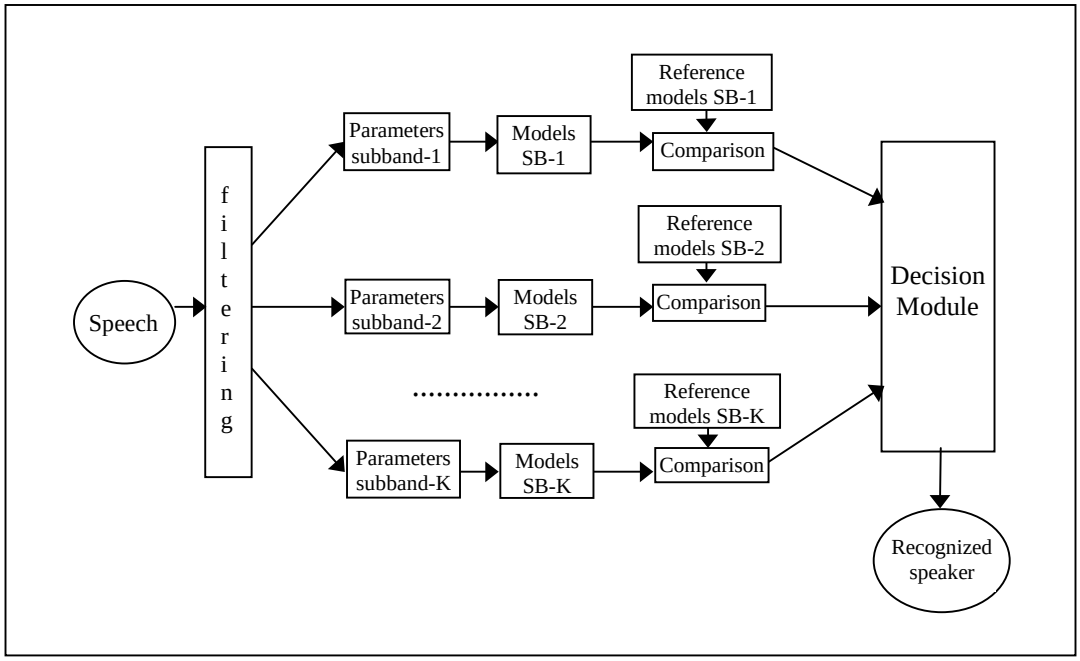

Figure 1 : block diagram of the multiband system

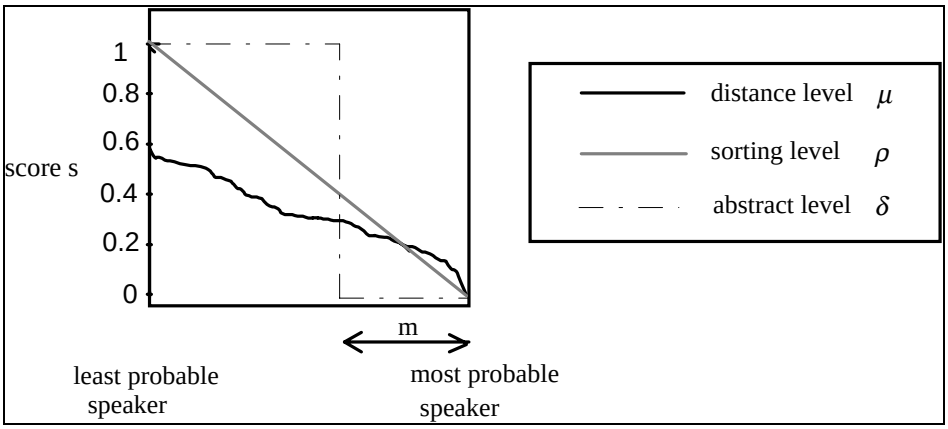

Figure 2 : different information levels at the output of each recognizer

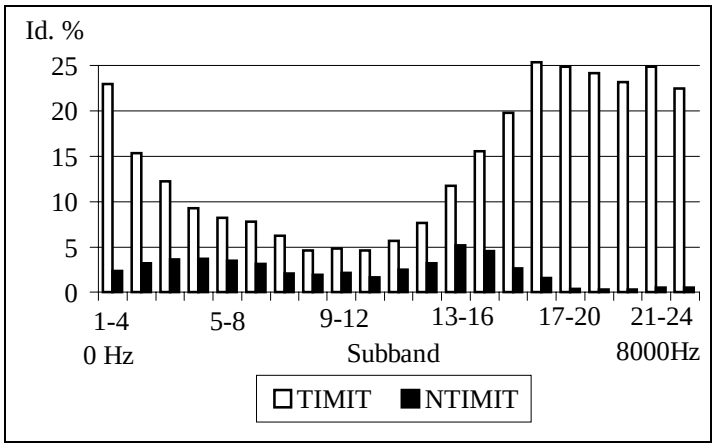

Figure 3 : isolated subband identification rates on TIMIT and NTIMIT (6s training/3s test - 630 speakers - 2925 tests)

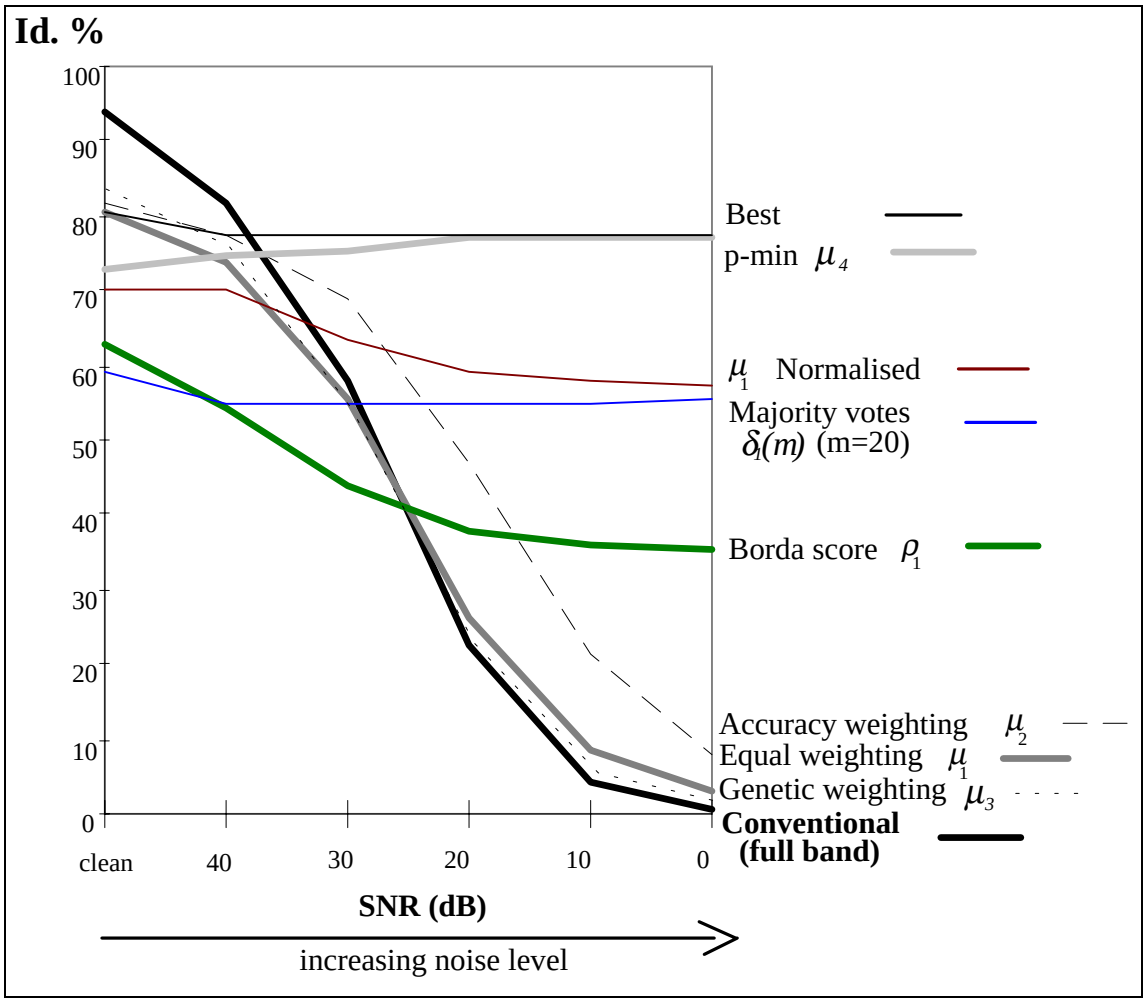

Figure 4 : speaker identification performances of different merging techniques on speech corrupted with bandpass limited noise at different SNRs (TIMIT_NBN - 6s Training/3s Test - 630 speakers - 2295 tests - architecture 6x4)

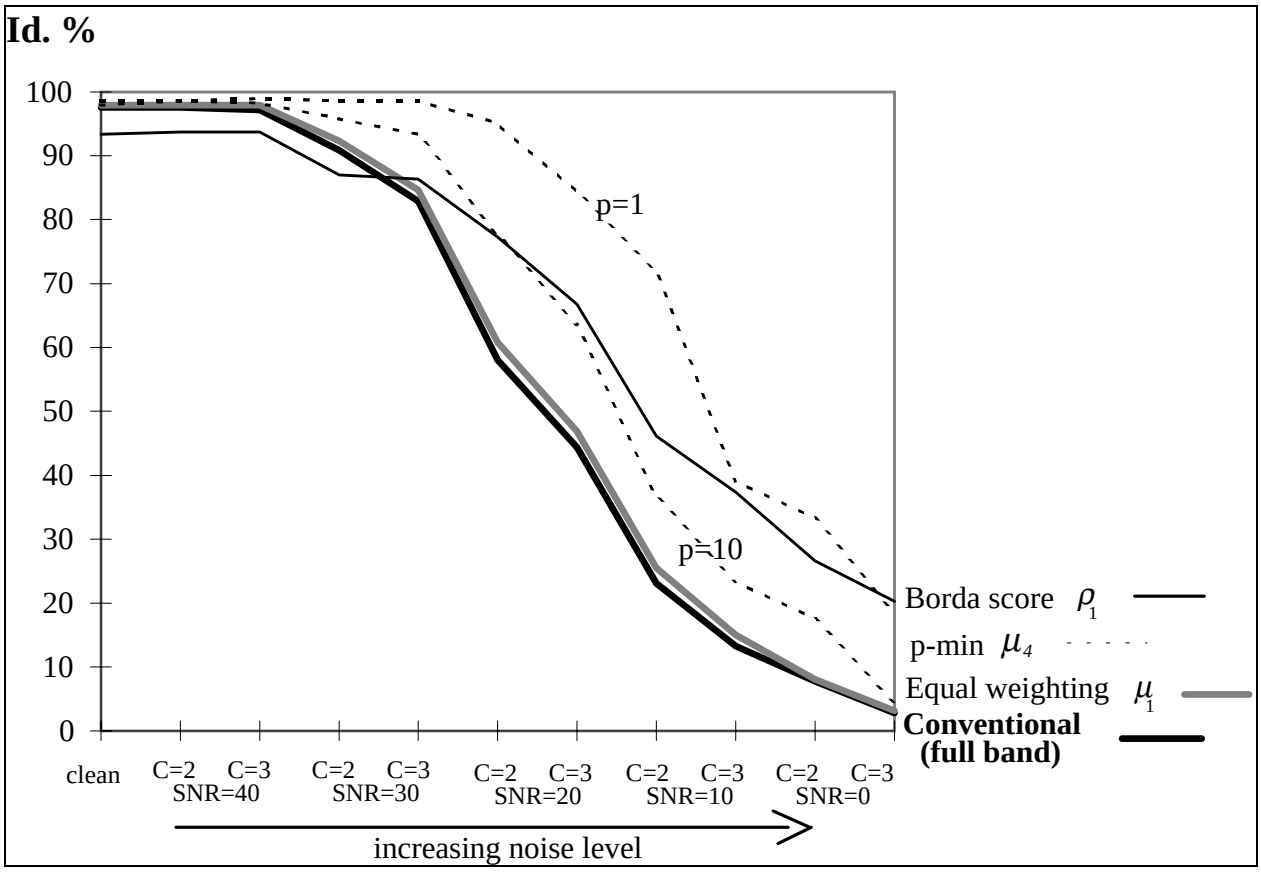

Figure 5 : speaker identification performances of different merging techniques on speech corrupted by an unpredictable noise randomly distributed on the whole spectral domain (TIMIT_FBN - 6s Training/3s Test - 63 speakers - 286 tests - architecture 24x20) - C : number of corrupted channels.

List of Tables

	TIMIT (630 locuteurs)	NTIMIT (630 locuteurs)
« Full sentences » protocol (~5.7s training / ~3.2s test) - 3150 tests	89.7 % id.	11.6 % id. (channels 1 to 17)
« Same duration » protocol (6s training / 3s test exactly) - 2925 tests	93.7 % id.	13.6 % id. (channels 1 to 17) 16.7% id. (channels 3 to 17)

Table 1 : comparison of performances (% identification) between the common protocol and our protocol on TIMIT and NTIMIT (same acoustic analysis for both protocols)

	Independent Subbands				Recomb. id%
channels	1-2-3-4-5-6	7-8-9-10-11-12	13-14-15-16-17-18	19-20-21-22-23-24	consecutive
id %	30.9	8.5	33.1	48.0	85.2
channels	1-5-9-13-17-21	2-6-10-14-18-22	3-7-11-15-19-23	4-8-12-16-20-24	crisscrossed
id %	30.5	25.7	27.5	29.0	58.4

Table 2 : speaker identification results on independent subbands and after recombination for a 4x6 architecture - consecutive or crisscrossed channels (TIMIT - 6s training/3s test - 630 speakers - 2925 tests)

number of subbands K	1	2	3	4	6	8	12
dimension of subbands *d*	24	12	8	6	4	3	2
total number of parameters used $K*(d^2+d)$	600	312	216	168	120	96	72
identification rates % (consecutive)	93.7	90.6	88.3	85.2	80.5	73.1	58.7
identification rates % (crisscrossed)	93.7	83.5	71.6	58.4	41.2	29.0	24.1

Table 3 : speaker identification results obtained after subband recombination for different divisions of the frequency domain (TIMIT - 6s training/3s test - 630 speakers - 2925 tests)

conventional 1x24 (630 speakers)	multiband 24x20 (630 speakers)	conventional 1x15 (630 speakers)	multiband 15x11 (630 speakers)
TIMIT - 6s Training / 3s Test		NTIMIT - 6s Training / 3s Test	
93.7 % id.	98.7% id.	16.7% id.	34.9% id.

Table 4 : potentiality of the proposed subband approach on TIMIT and NTIMIT (2925 tests)

conventional (full-band)	p-min. (p=1) μ_4	p-min. (p=10) μ_4	equal weighting μ_1	borda ρ_1
40.55	25.87	42.30	37.76	27.27

Table 5 : speaker identification results of different merging techniques (NTIMIT - 63 speakers - 6s training/3s test - 286 tests - architecture 15x11).