Localization and Selection of Speaker Specific Information with Statistical Modeling

L Besacier, J.F. Bonastre, C. Fredouille

Localization and Selection of Speaker Specific Information with Statistical Modeling

L. Besacier, J.F. Bonastre, C. Fredouille

Laboratoire Informatique Avignon

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

Abstract

Statistical modeling of the speech signal has been widely used in speaker recognition. The performance obtained with this type of modeling is excellent in laboratories but decreases dramatically for telephone or noisy speech. Moreover, it is difficult to know which piece of information is taken into account by the system. In order to solve this problem and to improve the current systems, a better understanding of the nature of the information used by statistical methods is needed. This knowledge should allow to select only the relevant information or to add new sources of information.

The first part of this paper presents experiments that aim at localizing the most useful acoustic events for speaker recognition. The relation between the discriminant ability and the speech's events nature is studied. Particularly, the phonetic content, the signal stability and the frequency domain are explored. Finally, the potential of dynamic information contained in the relation between a frame and its *p* neighbours is investigated.

In the second part, the authors suggest a new selection procedure designed to select the pertinent features. Conventional feature selection techniques (ascendant selection, knock-out) allow only global and *a posteriori* knowledge about the relevance of an information source. However, some speech clusters may be very efficient to recognize a particular speaker, whereas they can be non informative for another one. Moreover, some information classes may be corrupted or even missing for particular recording conditions. This necessity for

speaker specific processing and for adaptability to the environment (with no *a priori* knowledge of the degradation affecting the signal) leads the authors to propose a system that automatically selects the most discriminant parts of a speech utterance.

The proposed architecture divides the signal into different time-frequency blocks. The likelihood is calculated after dynamically selecting the most useful blocks. This information selection leads to a significative error rate reduction (up to 41% of relative error rate decrease on TIMIT) for short training and test durations. Finally, experiments in the case of simulated noise degradation show that this approach is a very efficient way to deal with partially corrupted speech.

Résumé

La modélisation statistique du signal de parole a été largement utilisée en reconnaissance automatique du locuteur. Les performances obtenues avec cette approche sont excellentes, en laboratoire. Cependant, une dégradation significative des performances est observée avec de la parole de qualité téléphonique ou bruitée. Pour palier ce problème, il est nécessaire de mieux comprendre la nature de l'information spécifique du locuteur exploitée par ces méthodes statistiques. Cette connaissance doit permettre de mieux prendre en compte l'information pertinente et/ou de mettre à contribution de nouvelles sources d'information.

La première partie de cet article reporte des expériences visant à spécifier les événements acoustiques les plus utiles à la reconnaissance du locuteur. Les liens entre le contenu phonétique du message, l'emplacement fréquentiel des informations, la stabilité du signal et les capacités de discrimination du locuteur sont successivement explorés. Enfin, la possibilité d'exploiter l'information dynamique contenue dans la relation entre une trame et les p suivantes est évaluée.

Dans une seconde partie, les auteurs proposent une nouvelle procédure de sélection de l'information spécifique du locuteur. En effet, les méthodes conventionnelles de sélection de paramètres (sélection ascendante, méthode du knock-out) ne permettent d'évaluer la pertinence d'une source d'information que de façon globale et *a posteriori*. Cependant, certains locuteurs sont mieux caractérisés par une source d'information que d'autres. De plus, la pertinence des sources d'information dépend de la qualité de l'échantillon de test. Face à ce besoin de traitements spécifiques suivant le locuteur et d'adaptation à l'environnement, nous proposons un système permettant de sélectionner automatiquement les parties les plus discriminantes d'une portion de parole.

L'architecture proposée divise le signal de test en blocs temps-fréquence. Le score de vraisemblance correspondant est calculé en sélectionnant dynamiquement les blocs temps-fréquence les plus pertinents. Une réduction significative du taux de mauvaise identification (jusqu'à 41% de réduction relative du taux de mauvaise identification sur TIMIT) est observée. Finalement, des expériences réalisées dans le cas d'un bruit simulé, montrent le potentiel de cette méthode pour traiter des signaux de parole partiellement dégradés.

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
$\{y_i\}_{tT+1\le i\le tT+T}$	t-th segment (of T frames) extracted from the speech sequence $\{y_t\}_{1\le t\le N}$
l_t	likelihood of acoustic vector y_t
l_t^k	likelihood of acoustic vector y_t on the k-th subband
L_t	average log-likelihood of the t-th segment
L_t^k	average log-likelihood of the t-th segment and of the k-th subband
H_t	normalized score of the t-th segment (homogeneous to a minus log-likelihood ratio)
H_t^k	normalized score of the t-th segment and of the k-th subband
$\mu(X,Y)$	similarity measure between speaker X and speaker Y
$\mu^k(X,Y)$	similarity measure between speaker X and speaker Y on the k-th subband
$Dev(y_t)$	stability criterion of acoustic vector y_t (frame t)
y_t^i	i-th component of acoustic vector y_t (frame t)

Pages #37

Tables #7

Figures #8

Keywords speaker recognition, speaker specific information, on-line selection, pruning, statistical modeling, time-frequency architecture

1. INTRODUCTION

A speaker recognition process can be basically divided into two main tasks. First, a speaker model is built from speech samples pronounced by a given person. Secondly, the probability that a speech recording corresponds to a given model is estimated and the final decision is made using this probability as well as information available *a priori*. Although this decision step is crucial for the performance of the system (many papers deal with the subject [9] [15] [18]), speaker specific information liable to influence the decision is also intrinsically contained in the speaker models. Therefore, it is worth trying to understand the nature of this information.

Statistical models are mainly used in speaker recognition. Most of them are based on the Hidden Markov Model (HMM) formalism. The different approaches can be derived with an increasing reduction in the number of models, the number of states per model and the number of gaussian densities per state. In that way, statistical approaches can move from large vocabulary continuous speech recognition-based models (LVCSR) towards monogaussian models (MGMs), which are made of a single state with only one gaussian. Gaussian mixture models (GMMs) seem to be an excellent compromise between performance and complexity and lead to the best recognition rates in text independent mode [25].

The speaker identification performance obtained with MGMs remains comparable to the one obtained with more complex models, like GMMs, for short training and test durations [6]. However, these basic models seem less efficient than GMMs for longer durations. Moreover, results obtained with statistical methods deteriorate dramatically for telephone speech (*Table 1*) or speech corrupted by noise (*Figure 1*).

Table 1.

In [6], experiments on TIMIT, FTIMIT (a restricted telephone bandwidth version of TIMIT) and NTIMIT (real telephone quality) have shown that bandlimiting is not the only problem for telephone speech; noisy environment as well as the difference between training and testing conditions in transmission channels and handsets are also a factor of degradation [28].

The fact that many environmental factors play an important role in the performance of a system shows that the information used may not be as speaker specific as expected. For instance, the microphone, the channel and the recording conditions influence the final decision significantly.

Figure 1.

In this work, the authors study the nature of the speaker specific information used by the models. This knowledge should allow to select only the useful information conveyed by the speech signal or to add new sources of information. MGMs, which are easy to implement and computationally efficient, are used. Moreover, the experiments will be performed with very short training and, as explained previously, more complex models cannot be implemented since not enough speech material is available to learn them correctly. However, it can be reasonably supposed that speaker specific information captured by an MGM will be caught by more complex models too. Conversely, a gain obtained with an MGM will not be systematically significant with more complex models.

The first part of this paper reports experiments aimed at localizing the most useful acoustic events for speaker recognition. These events differ, among other things, in their position in the time-frequency domain. At the temporal level, a former study on the discriminant ability of different phonemes is reported. Investigations aim at determining whether the most speaker specific information is rather situated in the transitions between phonemes or in the phoneme stable zones (targets). At the frequency level, speaker identification tests are conducted

independently on different subbands to know which part of the frequency domain is the most speaker specific. Finally, we propose to exploit the dynamic information contained in the relation between a frame and its *p* neighbours.

In the second part, the authors suggest a new selection procedure to deal with the redundancy observed between the various classes of information. Conventional feature selection techniques (ascendant selection, Knock-out [26]) allow only global and approximate knowledge about the relevance of an information source. However, the relevance of speech cues is speaker dependent rather than absolute [24], i.e. some speech clusters may be very useful to recognize a particular speaker, whereas they can be non informative for another one. Moreover, some parts of the information may be corrupted or even missing for particular recording conditions.

Thus, a new selection procedure is proposed to perform speaker specific processing and allow adaptability to changing acoustic environments. The most discriminant parts of a speech utterance are selected “on-line” with a maximum likelihood criterion, whereas the least informative parts are eliminated (pruning).

Section 2 describes the experimental conditions. *Section 3* is dedicated to the study of the speaker specific information used by these models. In *Section 4*, the “on-line” selection method is detailed and then experimented with for the special case of a time-frequency architecture in *Section 5*. *Section 6* concludes this work and shows that the proposed approach allows interesting feedback on the localization of speaker specific information when an *a posteriori* analysis of the rejected speech parts is performed.

2. REFERENCE SYSTEM

2.1 Monogaussian Modeling of Speakers

The monogaussian modeling is the starting point of the proposed system. It is more precisely described in [6].

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 mean vector $\bar{x}$ and covariance matrix X:

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

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:

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

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

$$\bar{L} = \frac{1}{N}\sum_{t=1}^{N} \log(l_t) \quad (3)$$

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

$$\mu(X,Y) = \mu(X, y_1^N) = -\bar{L} \quad (4)$$

This measure is equivalent to the standard gaussian likelihood measure (asymmetric μ_G) defined in [6]. The following symmetric version of this measure (β symmetrisation [6]) is defined as:

$$\mu_{G_{[\beta_{MN}]}}(X,Y) = \frac{M.\mu(X,Y) + N.\mu(Y,X)}{M+N} \qquad (5)$$

The symmetric version of the measure is used since it is shown in [6] that symmetrisation has a positive effect when little speech material is available (up to 30% of error reduction with short training and testing).

2.2 Experimental Conditions

2.2.1 Databases

TIMIT and NTIMIT databases are used during the various experiments. Even if these databases are mono session, they offer the advantages of being largely used in the literature for comparison, being suited to text independent task, and proposing a large number of speakers.

TIMIT database [10] contains 630 speakers (438 male and 192 female speakers), 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).

The NTIMIT database [17] was obtained by playing TIMIT speech signal through an artificial mouth installed in front of the microphone of a fixed handset and by 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).

2.2.2 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. For each frame, spectral vectors of 24 Mel-Scale Triangular-Filter Bank coefficients (24 channels) are calculated from the Fourier Transform power spectrum and expressed in logarithmic scale[1]. 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 [2] [3] [4] [5] [6].

Finally, it can be noticed that a subset of filterbank coefficients can be directly interpreted as a frequency subband. Thus, speaker identification experiments on independent subbands can be conducted easily.

2.2.3 Training and Test Protocols

In the proposed 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 taken into account. 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 test blocks per speaker).

The reference and test patterns are thus computed from exactly the same number of samples for each speaker. These exactly identical durations were required only for the pruning

experiments reported in *Section 5*; however, this new protocol is used in all experiments presented in this paper. This protocol yields results comparable to those obtained with the regular protocol used on TIMIT ("phrase by phrase" protocol) [5].

All the tests are done within the framework of text-independent closed-set speaker identification.

3. TRACKING SPEAKER SPECIFIC INFORMATION

3.1 Phonemes

In [19], the authors observed that the speaker identification performance (obtained with MGMs) changes according to the phonetic label of the speech segments used. These results tend to show that the speaker dependent information captured by MGMs is consistently common to all phonetic classes and that the phonetic homogeneity of the test material may improve the quality of the estimates. Thus, speaker specific information extracted with the MGMs is not equally distributed in the speech signal. A large redundancy in the information conveyed by the different phonemes is observed. All classes of phonemes give good results alone. Therefore, an intelligent use and selection of these different sources of information should authorize significant performance enhancement.

3.2 Stability

The studies reported in the previous section show that the use of phonetically homogeneous segments improves performance. Two hypotheses, which are not conflicting, can explain this result. On the one hand, the phonetic content of segments is important. This is confirmed by the difference in performance, observed in [7] [14] [19] [20], between various phonetic segments. On the other hand, the homogeneity of segments can also contribute to increasing performance since the modeling used is based on statistical methods. Indeed, this kind of methods determines the relevance of a piece of information from its repetitive nature. In the

case of this study (gaussian mixture-based models relying on spectral feature vectors), it can be reasonably assumed that very unstable zones of speech signal, such as transitions between phonemes, may be less finely modeled than stable zones, independently of the amount of speaker specific information initially present in the speech signal.

Consequently, this section aims at determining whether the selection of stable zones of the speech signal, which correspond mainly to phoneme kernels (but also to silence zones and occlusions), can lead to performance improvement.

Therefore, two kinds of experiments have been conducted. The first one studies the global performance of the identification system according to the quantity of "stable" zones used. The second one has to demonstrate a possible correlation between the stability level of a test frame uttered by a given speaker X and the likelihood estimation between this frame and the speaker model.

In these two contexts, the same stability criterion, which allows to assign a stability coefficient to each frame, is used. The criterion is based on the behavior of frame *t* compared to the one of *(N/2-1)* frames around it. In practice, assuming that a mean spectrum is computed from an *N* frame time window centered on frame *t*, stability criterion $Dev(y_t)$, defined in *(6)*, is the distance between *p*-dimensional vector y_t associated with frame *t* and that mean spectrum represented by *p*-dimensional vector $\bar{y}$:

$$Dev(y_t) = \frac{1}{p} \sum_{i=1}^{p} (y_t^i - \overline{y^i})^2 \qquad (6)$$

with $\bar{y} = \frac{1}{2N+1} \sum_{t=1}^{2N+1} y_t$ (7)

3.2.1 Experiments

The first experiment consists in reporting the identification rates obtained according to the amount of the most stable frames selected during test and according to the training protocol used. Two kinds of training protocols are proposed:

1. a *classical training* during which all the training data are used: 6s of speech signal corresponding to 600 frames.
2. a *"stable" training* during which the models are estimated by using a reduced amount of training data composed of the 300 or 500 most stable frames selected among the 600 initial training frames.

In order to evaluate the potential of stable frame selection, similar experiments have been conducted by selecting frames of speech signal randomly.

Table 2.

Table 2 reports identification rates obtained by using one of the three training types: classical, "stable" or random (300 or 500 stable/random frames selected) and by selecting 50, 150, 300 stable/random frames during testing. The pair "Classical training (600)/300 frame-based testing" is considered as the reference system.

Different remarks can be made:

- With classical training, no gain is observed by selecting the most stable zones during test, compared to results obtained with random zones.
- Similarly, selecting stable zones during training does not improve performance even if the selection of the most stable zones is also applied during test. Besides, results are biased by the reduction in training data - mainly observed with the 300-stable-frame-based training protocol -, which involves a dramatic decrease in performance due to a bad estimate of models.

- Selecting unstable zones during test (in the same training conditions) gives worse identification rates which are not provided here.

During the second experiment, still based on identification test and on classical training, the numerical pair, *stability coefficient and likelihood*[2], is computed for each test frame. The estimate of the correlation rate between the two distributions, *stability coefficient* vs. *Likelihood,* leads to a mean result of -0.06. Therefore, no apparent correlation exists between the stability level of a frame and its discriminant power.

3.2.2 Conclusion

As observed during the previous experiments, stable zones of the speech signal do not seem to convey more specific information than zones selected randomly. This tends to confirm that the phonetic nature of speech segments (both test and training segments) is more important for speaker characterization than the homogeneity of segments. But it is important to bear in mind that stable zones also include silence and occlusion parts of the speech samples.

3.3 Frequency Subbands

3.3.1 Subband Modeling

The following 'K-subband' model of speaker X can be obtained from the initial full-band model:

$$M_X(K) = \left\{\left(X^1, \bar{x}^1\right),\ldots,\left(X^k, \bar{x}^k\right),\ldots,\left(X^K, \bar{x}^K\right)\right\} \qquad (8)$$

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 covariance matrix X and $\bar{x}^k$ is a sub-vector of mean vector $\bar{x}$ (X and $\bar{x}$ being computed on the whole spectral domain).

Therefore, the quantities defined in *(2) (3)* and *(4)* can be respectively written for the k-th subband:

- l_t^k likelihood of acoustic vector y_t on the k-th subband,

- $\overline{L}^k$ average log-likelihood of $\{y_t\}_{1\le t\le N}$ on the k-th subband,

- $\mu^k(X,Y)$ similarity measure between speaker X and speaker Y on the k-th subband.

Figure 2.

3.3.2 Experiments on Isolated Subbands

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 one defined in *(5)* and applied to each subband.

Figure 2 shows the speaker identification performance obtained on each isolated subband for 6s training/3s test on TIMIT and NTIMIT databases.

Large differences between subbands are observed, which shows that speaker specific information is not equally distributed on the spectral domain. 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 sharp 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 telephone network noise and to signal distortions.

3.3.3 Channel Selection

A channel selection method is proposed to estimate more precisely the relative effectiveness of each part of the frequency domain. The method used is the ‘knock-out’ procedure [26]. The method begins by evaluating the effectiveness of each of the N=24 channel subsets composed of N-1 channels. The most effective subset is then determined and the channel not included in

this subset is defined as the least important channel. This channel is then eliminated (or 'knocked-out') and the descending procedure continues until all the channels are 'knocked-out' from consideration.

Table 3 shows the speaker identification rates obtained with the best set of channels on TIMIT (630 speakers) compared to the full-band results. The results obtained with half of the channels and with channels representing half of the frequency domain are also reported in this table.

Table 3.

The best identification results are obtained with 18 channels[3] on TIMIT (94.3%) corresponding to 80% of the whole frequency domain. These results represent a slight error rate reduction compared to the same full-band test (93.7%). However, this improvement may be only considered as an *a-posteriori* optimization of the results on the current database.

Good performance is still obtained when using only half of the channels[4]: 89.5% identification rate on TIMIT for 12 well-chosen channels; the main part of the speaker specific information is thus condensed in about 60% of the total frequency domain.

3.4 Dynamic Information

Many studies have been dedicated to the exploitation of dynamic information in speaker recognition systems [1] [13] [21] [23] [27]. They have shown the interest of this kind of data as another source of information, since they obtain performance similar to static information one and they are more robust in noisy environments.

Various approaches are proposed in the literature to exploit dynamic information: extraction of derivatives of the function time of instantaneous features during the parameterization (Delta and Delta-Delta coefficients), use of predictive models or static methods applied to dynamic information… However, these methods do not allow to fully exploit dynamic

information on a large time window without involving some computation complexity problems or requiring a too large amount of data to train models.

This study aims at considering a sufficient time window (100ms of speech signal) to exploit dynamic information in depth by using methods suited to cope with the previous problems (training data and processing complexity).

3.4.1 "Dynamic" Modeling

The method proposed here is based on statistical methods applied to dynamic vectors stemming from the concatenation of *T* consecutive frames of speech signal [11]. This method is associated with the multi-band approach in order to significantly reduce computation complexity problems. Indeed, in practice, a full band approach will lead to consider dynamic vectors of 240 coefficients if a time window of 10 successive frames (parameterized by 24-dimensional vectors) is considered, whereas a multi band approach, based on 6 subbands, leads to process individual dynamic vectors of 40 coefficients each.

3.4.2 Dynamic experiments and results

Table 4 (third column) provides the identification rates of experiments conducted on dynamic subbands presented in the previous section. Results obtained on static subbands are also given for comparison.

It can be observed that results differ between the subbands. Dynamic subbands: 13-16, 17-20 and 21-24 show a slight performance improvement, whereas dynamic subbands: 1-4, 5-8 and 9-12 lead to performance decrease if compared to identification rates obtained by static subbands.

This may be due to the concatenation of successive frames of speech signal (required to exploit dynamic information), which involves taking a large amount of features into account and leads to great information redundancy.

The selection of useful dynamic information can be a solution to cope with this problem. The next section presents the methods used and the results obtained after selection.

3.4.3 Selection of dynamic coefficients and results

The goal of the selection procedure is to extract an optimum subset from a set of dynamic coefficients, (associated with an individual subband), which will lead to enhance the identification system. In this perspective, an ascendant method [8] (variant of the knock-out selection procedure [26]) associated with a selection criterion based on the identification rate is applied on each subband [11].

Experiments for the selection of optimum subsets[5] are conducted on a development set of 135 tests (stemming from the first 63 males of TIMIT) and the fourth column of *Table 4* gives identification rates obtained using these same optimum subsets applied to a second set of 2639 tests (stemming from speakers of TIMIT without the previous 63 speakers). During training, the 630 speakers of TIMIT are used.

These results show, on each individual subband, a significant performance improvement compared to dynamic subbands without selection and to static ones.

This highlights the necessity for a selection of the useful information and demonstrates the potential of dynamic information to characterize speaker in this selection context.

Table 4.

3.4.4 Recombination of dynamic subbands

The final step of a multiband approach consists in recombining individual measures obtained on each subband in order to yield a final decision for the recognition task.

This fusion step is applied to the dynamic subband measures obtained with the selection of the best coefficients as seen in the previous section. A basic arithmetic mean is chosen in order to

recombine dynamic subband results since the main interest of this fusion is to demonstrate the potential of dynamic subbands and not to optimize the system.

Table 5, which provides recombination results of dynamic subbands, shows a significant degradation of performance if compared to recombination results of static subbands.

The increase in performance observed individually on each dynamic subband does not allow to improve the global recognition decision. This could be explained by:

- a large redundancy of information between the different subbands
- the necessity for a more "clever" recombination method (than a simple arithmetic mean). For example, a frame-based recombination (with a weight according to the dynamic nature of the block) should yield better results.
- and a coefficient selection criterion more suitable for the recombination step.

Table 5.

3.4.5 Application in the NIST 99 speaker recognition evaluation campaign

The dynamic approach, coupled with the selection of relevant features, as described in the previous section, was implemented in the framework of speaker verification and evaluated during the NIST 99 speaker recognition evaluation campaign.

In this evaluation context, the database was made up of speech signal recordings issued from Switchboard database and built from concatenated telephone conversation segments.

Three different recognizer schemes were evaluated:

- SFB referring to a simple static Full Band;
- SFB+DSB composed of a Static Full Band associated with three Dynamic SubBands;
- DFB+DSB composed of a Dynamic Full Band and three Dynamic SubBands.

Both of them were quite different from the system baseline described in this paper. First, they consisted of cepstrum parameter-based recognizers. Besides, EM trained GMMs were used to

model each speaker since sufficient speech material (more than one minute) is available for the training (a 16 gaussian mixture summarized by full covariance matrix for the static full band and a 128 gaussian mixture summarized by diagonal covariance matrix for the dynamic full band and subbands).

Finally, it has to be noticed that pertinent feature selection was used for the dynamic subbands only and was carried out through an MGM-based identification system, using a separate data set (extracted from the NIST 98 evaluation campaign data).

Figure 3 provides a comparison, in terms of DET curves, between the three different recognizer architectures presented above: SFB, SFB+DSB, and DFB+DSB.

It can be observed that both dynamic and static recognizers (SFB+DSB) lead to some performance improvement if compared to the static recognizer alone (SFB). On the other hand, the fully dynamic system (DFB+DSB) outperforms the two others.

These results highlight the well-known robustness of dynamic information [13][27] in a telephone and noisy environment.

Nevertheless, they do not demonstrate the gain in terms of performance involved by the feature selection. This last point should be investigated in future work.

Figure 3.

4. SELECTION OF THE SPEECH SEGMENTS

4.1 Motivation

In the previous section, conventional procedures are used to select the most useful spectral information (*section 3.3.3*) or to deal with the redundancy induced by a dynamic approach

(*section 3.4*). However, it seems that these techniques (ascendant selection, knock-out method) allow only a global and approximate knowledge about the relevance of a set of features. Indeed, some speech clusters may be very useful to recognize a particular speaker, whereas they can be non informative for another one. Moreover, according to the recording conditions, some types of information may be corrupted or even missing.

To cope with these various problems, a system which dynamically selects the best speech parts of a test utterance, according to the speaker model concerned, is proposed.

4.2 "On-line" Selection with Maximum Likelihood Criterion

4.2.1 Principle

Several likelihood scores $(S_i(X))_{i=1..I}$ can be calculated from the different parts of a given test utterance, compared to the model of speaker X. Instead of averaging these scores, some of them are eliminated (pruning) and the final decision is made with a limited number of partial scores.

The likelihood scores correspond to different events and they must be first normalized in order to make comparison between them meaningful. In fact, if the likelihood score of a speech part is lower than the likelihood score of another one, it does not necessarily mean that the first speech part is less informative than the second one, because both parts convey different information and there is no basis for a meaningful comparison between them.

Consequently, a log-likelihood ratio is used as a normalized score, as defined in [16]:

$$\log S_i norm = \log S_i(X) - \max_{Z \neq X} \log S_i(Z) \qquad (9)$$

In the experiments, normalizing speakers Z, for a given person X, will be chosen among the other reference speakers rather than among a completely separate group. Speakers are thus normalized by each other.

Then, the pruning process is based on the assumption that the maximum likelihood scores resulting in correct identifications are in general higher than the maximum likelihood scores resulting in inaccurate identifications. In other words, when a part of the speech signal is error-prone (i.e. when the true speaker is not identified on this particular speech event), it is not due to a non-target speaker model matching the speech part well, but rather to the true speaker model performing badly.

For convenience, a minus-log-likelihood ratio H_i, which is equivalent to a distance measure, is used:

$$H_i = -log\, S_i norm \qquad (10)$$

H_i is also called discriminant function [12] (p.52) since if $H_i < 0$, speaker X scores higher than everyone else on the given speech part and so speaker X is recognized on the single part; if $H_i > 0$, the speaker recognized on this part is not speaker X.

4.2.2 Potential of the ML Criterion

The potential of the maximum likelihood criterion is illustrated by *Figure 4* where the distributions of normalized frame scores H_t (here, 1 speech part=1 frame) are represented for speaker models which score higher than everyone else on a given frame (i.e. negative values of H_t). Two types of frames are distinguished: frames on which the target speaker would be recognized if the decision was made on a single frame (successful frames) and frames on which a non-target speaker would be recognized (unsuccessful frames). The distributions of both classes are equivalent to the density functions of H_t and can be noted respectively $p_H(H/X)$ and $p_H(H/\overline{X})$.

It can be observed that the frames may have lower minus log-likelihood ratios H_t for the true speaker ($p_H(H/X)$) than for non-target speakers ($p_H(H/\bar{X})$), which tends to prove the need for a pruning process to select the lowest values of H_t and thus eliminate error-prone frame scores.

Figure 4.

5. PRUNING EXPERIMENTS

5.1 Block-based Architecture

The results obtained in *section 3* have shown that speaker specific information is not equally distributed both at the temporal and frequency levels. However, instead of using an analytical approach to extract the different speech parts at the input of the selection system, an arbitrary division of a speech utterance into several time-frequency blocks has been chosen.

A test utterance is thus split into 'n' time segments and into 'K' frequency subbands (*Figure 5*), with a possible overlap between subbands. For each pair (t,k), corresponding to segment 't' and subband 'k', an average log-likelihood score $\bar{L}_t^k$ can be calculated:

$$\bar{L}_t^k = \frac{1}{T}\sum_{i=1}^{T} log(l_{tT+i}^k) \qquad (11)$$

Figure 5.

A log-likelihood ratio is then used as a normalized score, as defined in *(9)*:

$$\bar{L}_t^k norm = L_t^k(X) - \max_{Z \neq X} L_t^k(Z) \qquad (12)$$

and the minus-log-likelihood ratio $\bar{H}_t^k$, equivalent to a distance measure, becomes:

$$\bar{H}_t^k = -\bar{L}_t^k norm \qquad (13)$$

Pruning is then achieved on these normalized scores; the final score is:

$$\bar{H} = \min_{p,q} \sum_{1}^{p}\sum_{1}^{q} \bar{H}_t^k \qquad (14)$$

The $p*q$ lowest block scores are averaged for each speaker, with $p<n$ (n number of segments in the test utterance) and $q<K$ (K number of subbands in the architecture) ; p and q do not have sence independently since the product pq specifies the selected blocks ratio.

Finally, two special cases can be derived from this general formalism:

- $K=1$ and $n>1$ correspond to a "segment level normalization approach" [22] and only time pruning is considered [3],

- $n=1$ and $K>1$ correspond to a "multiband approach" [2] and only frequency pruning is considered.

Note that the blocks selected in the sum can vary according to the speaker model considered.

5.2 Time Pruning

The special case $K=1$ (full-band model) is considered here. The influence on the performance of the number of selected (i.e not discarded) segments (p) is investigated when a segment is composed of a single frame (T=1). The results are reported in *Figure 6* (300 frames / test utterance).

For both databases, optimum results are obtained when some frames are pruned: id.=100% for p=150 on TIMIT and id.=43% for p=260 on NTIMIT. This shows that information selection is important since some frames in a test utterance can contaminate the final score. Moreover, it is interesting to note that a reasonably good performance is obtained on TIMIT when a single frame per speaker is kept (71.63% id.), i.e. when an extremely small amount of speech is used for each speaker to make the final decision ! (This part of speech signal being often different from one speaker to another).

Figure 6.

5.3 Time-frequency Pruning

An architecture of 24 subbands of 20 channels each (24x20) is experimented with for TIMIT and an architecture of 15 subbands of 11 channels each (15x11) for NTIMIT. The segment size is T=1 (i.e. 1 segment=1frame). For a 3s test duration (300 frames), the total number of time-frequency blocks is then 7200 on TIMIT and 4500 on NTIMIT. The influence of the number of blocks selected *pq* is investigated. The results are reported in *Figure 7*.

For both databases, the best results are obtained when some blocks are pruned: id.=100% for *pq*=3500 or 4500 on TIMIT and id.=41.95% for *pq*=3900 on NTIMIT. However, it is difficult to see the real benefit of the joint time and frequency pruning process in comparison with the single time-pruning technique.

Figure 7.

5.4 Validation

The best values of *p* (time pruning, *Section 5.2*) and *pq* (time-frequency pruning, *Section 5.3*) obtained for 63 speakers on TIMIT and NTIMIT are used to validate the benefit of the pruning procedure for speaker recognition. Speaker identification tests are conducted on the 567 remaining speakers of TIMIT and NTIMIT. The final test set is completely distinct from the tuning set from which the optimal values of *p* and *pq* are evaluated. The identification results obtained are presented in *Table 6*. For both databases, performance improvement is significant. The time-frequency pruning procedure leads to a 41% error rate reduction on TIMIT, compared with the conventional monogaussian classifier. However, the benefit of the joint time and frequency pruning procedure in comparison with the single time pruning process is less evident on NTIMIT.

Table 6.

5.5 Noisy environment

To evaluate the gain in terms of robustness, an experiment is conducted in a noisy environment. For this experiment, a noise was added to the TIMIT test signal. A simulated noise was chosen in order to make the experiment easy to reproduce.

Some remarks resume the protocol used:

- Both the model training and the meta parameter tuning of the system were performed on clean speech signal.
- No a priori knowledge about noise is used.
- The noise is unpredictable and distributed on the whole spectrum as follows: for each frame, *C* (*C*=2 or 3) frequency channels among 24 (dimension of the full-band acoustic vector) were randomly selected and degraded for different SNRs.

The potential of the pruning process is illustrated by results in *Table 7*. In every case, the time-frequency pruning approach widely outperforms the conventional one, which seems to be very promising. It can be noticed that, in this case, the models and the optimal number of blocks pruned (pq) were learned on clean speech material, which shows the adaptability of the pruning procedure without *a priori* knowledge on the degradation affecting the test signal. Obviously, this relative gain has to be confirmed in non-simulated (real) noisy conditions.

Table 7.

6. CONCLUSION

6.1 Summary

In this paper, the nature of the speaker specific information used by statistical models has been discussed. The various investigations demonstrate the difficulty in highlighting this information. For example, it can be intuitively supposed that the most stable zones of speech signal would be more finely modeled by statistical approaches and consequently should lead

to enhance system performance if they could be isolated. Nevertheless, experiments in this way do not show significant results. However, independently of performance, it is necessary to know the nature of the information classes used.

It has been shown that this information is not equally distributed according to the phonetic content of speech segments and in the time-frequency domain. A large redundancy is observed between the various classes of information in many cases (even though a more complex approach is proposed such as dynamic modeling). Therefore, this demonstrates the necessity of selecting the useful part of information conveyed by the speech signal.

The selection methods suggested in this paper (Knock-out, ascendant method…) enhance recognition performance. For example, the channel selection (see *section 3.3.3*) allows to reach an optimum identification rate of 94,3% (on TIMIT) by selecting only 83,2% of the frequency domain. But, these conventional selection techniques allow only global and approximate knowledge about the relevance of a set of features. Nevertheless, some speech clusters may prove very useful to recognize a particular speaker, whereas they can be non-informative for another one. Moreover, according to the recording conditions, some types of information may be corrupted or even missing.

We have proposed a system which selects the best parts of the speech signal dynamically during test according to the speaker model concerned. The selection was based on a maximum likelihood criterion.

This “on-line” selection procedure was experimented with for the special case of time-frequency architecture. The results obtained have shown that this technique can significantly increase the performance of a speaker identification system in normal or noisy (simulated) conditions. Nevertheless, further investigations have to be conducted in order to test the robustness of this approach against noise in real conditions.

6.2 Knowledge gathered from the result analysis

The originality of the proposed selection technique is that the speech parts selected on a same test utterance can vary from one speaker model to another. The analysis of the rejected (or selected) parts allows interesting feedback. This analysis was made for the particular case of time pruning (*section 5.2*). *Figure 8* shows the distribution of the frames according to their frequency of selection when the final score is computed with only half of the frames. In other words, if *Nselect*=63, the corresponding frames were used by all 63 speaker models; if *Nselect*=0, the corresponding frames were rejected by all speaker models, during the recognition stage.

Figure 8.

The profile of the results suggests two different conclusions:

- a coherence exists in the information conveyed by the frames, i.e. when a frame is rejected (or selected) by a speaker model, it is rejected (or selected) by the majority (parts 1 and 3),
- however, it is also clear that the speech frames selected are different from one speaker to another (part 2), which confirms that specific information is not the same according to the speaker concerned. The inadequacy of conventional selection processes is then clearly pointed out by this analysis.

6.3 Outlook

To go further, it would be interesting to know the phonetic label of the frames kept and the frames rejected. Performing a more systematic post analysis will allow to further investigate the phonetic aspect of speaker identification.

Finally, we also intend to apply the “on-line” selection method during the training phase, which should be an interesting approach to refine the speaker models.

7. REFERENCES

[1]C. Bernasconi, “On instantaneous and transitional spectral information for text-dependent

speaker verification". *Speech Communication,* 9(2), pp 129-139, April 1990.

[2]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.

[3]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.

[4]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.

[5]L. Besacier, "Un modèle parallèle pour la reconnaissance automatique du locuteur". PhD Thesis, University of Avignon, April 1998.

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

[7]J. F. Bonastre and H. Méloni, "Inter and intra-speaker variability of French phonemes: advantages of an explicit knowledge based approach". *In workshop on Automatic Speaker Recognition,* pp 157-160, Martigny, Switzerland, April 1994.

[8]D. Charlet and D. Jouvet, "Optimisation du jeu de paramètres acoustiques pour la vérification du locuteur". *XXIèmes Journée d'Etudes sur la Parole*, pp 399-402, Avignon, France, 1996.

[9]D. Charlet, D. Jouvet and O. Collin, "An alternative normalization scheme in HMM based text dependent speaker verification". *In Proc. on Speaker Recognition and its Commercial and Forensic Applications (RLA2C),* Avignon, France, April 1998.

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

[11]C. Fredouille and J.F. Bonastre, "Use of dynamic information with second order statistical methods in speaker identification". *In Proc. on Speaker Recognition and its Commercial and Forensic Applications (RLA2C),* Avignon, France, April 1998.

[12]K. Fukunaga, "*Statistical Pattern Recognition"*. Second Edition, Academic Press, Inc., San Diego, 1990.

[13]S. Furui, "Cepstral analysis for automatic speaker verification", *IEEE Transactions on ASSP*, vol. 29(2), pp 254-272, 1981.

[14]S. Furui, "An overview of speaker recognition technology". *In workshop on Automatic Speaker Recognition,* pp 1-9, Martigny, Switzerland, April 1994.

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

[16]H. Gish and M. Schmidt, "Text independent speaker identification". *IEEE Signal Processing Magazine*, pp 18-32, October 1994.

[17]C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, "NTIMIT: A Phonetically Balanced Continuous Speech, Telephone Bandwidth Speech Database". *In Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP),* April 1990.

[18]J. Lindberg, J. Koolwaaij, H.P. Hutter, D. Genoud, J.B. Pierrot, M. Blomberg and F. Bimbot, "Techniques for a priori decision threshold estimation in speaker verification". *In*

Proc. on Speaker Recogni[6]*tion and its Commercial and Forensic Applications (RLA2C)*, Avignon, France, April 1998.

[19]I. Magrin-Chagnoleau, J.F. Bonastre and F. Bimbot, "Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods". *In Proc. Eurospeech-95*, Madrid, Spain, September 1995.

[20]O. Mella, "Pertinence des trois premiers formants des voyelles orales dans la caractérisation du locuteur". *XIXèmes Journée d'Etudes sur la Parole*, pp 549-554, Brussels, Belgium, 1992.

[21]I. Magrin-Chagnoleau, **"**Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte", PhD Thesis, Ecole Nationale Supérieure des Télécommunications, Paris, 1997.

[22]K. Markov and S. Nakagawa, "Frame level likelihood normalization for text-independent speaker identification using GMMs". *In Proc. ICSLP 96*, pp 1764-1767, Philadelphia, USA, 1996.

[23]C. Montacié and J.L. Le Floch, "Discriminant AR-Vector models for free text speaker verification". *In Proc. Eurospeech-93*, pp 161-164, Berlin, Germany, September 1993.

[24] F. Nolan, "*The phonetic bases of speaker recognition"*. CUP 1983. Cambridge.

[25] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models". *Speech Communication*, vol. 17, pp 91-108, August 1995.

[26] M.R. Sambur, "Selection of acoustic features for speaker identification". *In IEEE Transactions on ASSP*. n°23(2), pp 176-182, April 1975.

[27]F.K. Soong and A.E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition". *IEEE Transactions on ASSP*, vol. 36(6), pp 871-879, June 1988.

[28] S. Van-Vuuren, "Comparison of Text independent speaker recognition methods on telephone speech with acoustic mismatch". *In Proc. ICSLP 96*, pp 1788-1791, Philadelphia, USA, 1996.

List of Tables.

TIMIT	NTIMIT
93.7 % id.	16.7% id.

1. comparison of speaker identification performance between TIMIT and NTIMIT databases (normal and telephone quality) – MGMs – 24 filterbank coeff. (TIMIT) or 17 (NTIMIT) - 6s training / 3s test - 630 speakers - 2925 tests - [5]

	Test					
	% Identification according to number of the most stable frames selected			% Identification according to number of random frames selected		
Training	50	150	300	50	150	300
Classical (600)	77.3	93	94.8	89.5	94.8	94.8
Stable (300)	72.4	76.9	67.5	60.1	62.9	67.4
Stable (500)	77.6	92.7	94.8	84.6	94.1	94.8
Random (300)	71.3	81.1	82.9	74.5	80.8	82.9
Random (500)	75.2	92.3	94.1	86.7	92.3	94.1

2. Identification rates (in %) obtained according to different types of training and various numbers of stable and random frames selected during test (630 speakers - 286 tests)

	FULL BAND	BEST RESULTS	HALF OF THE CHANNELS	HALF OF THE FREQ. DOMAIN
Number of channels	24	18	12	9
% of the full freq. domain	100.0%	83.2%	64.4%	50.0%
Id. %	93.7%	**94.3%**	89.5%	79.4%

3. Main identification results with the channel selection procedure - TIMIT (6s training/3s test - 630 speakers - 2925 tests)

	Static SB	Dynamic SB without Selection	Dynamic SB with Selection
SB	% Id.	% Id.	% Id.
1-4	21.4	19	23.9
5-8	8.3	6.4	8.8
9-12	4.9	4.1	5.3
13-16	10.8	11.4	12.6
17-20	24.3	24.6	27.8
21-24	22.2	25	26.6

4. Identification rates obtained by using static subbands, dynamic subbands without any selection (integrating the 40 coefficients) and dynamic subbands with selection of best coefficients (6s training/3s test - 567 speakers - 2639 tests).

Id. rate after recombining static subband measures	Id. rate after recombining dynamic subband measures
79.4	76.4

5. Identification rates obtained after recombining the measures of dynamic subbands and those of static subbands (6s training/3s test - 567 speakers - 2639 tests).

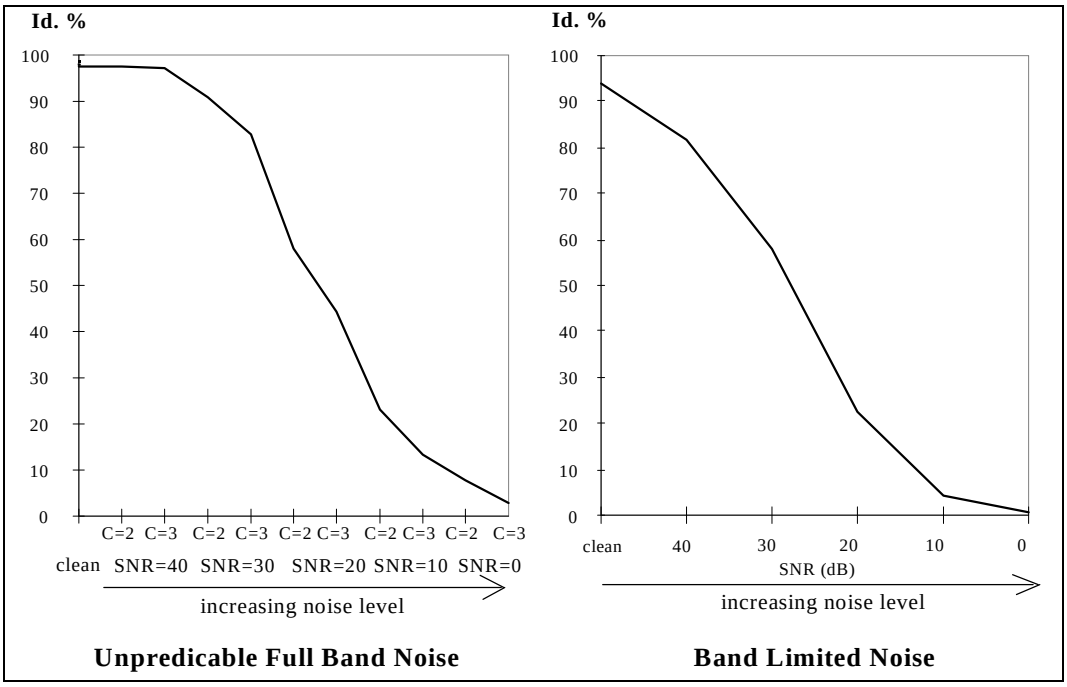
	BASELINE NO PRUNING	TIME PRUNING	TIME-FREQUENCY PRUNING
TIMIT	n=1; K=1	K=1; p=150; T=1	K=24; pq=4500; T=1
Id. %	91.66	94.20	95.14
NTIMIT	n=1; K=1	K=1;p=260;T=1	K=15; pq=3900; T=1
Id. %	15.91	18.64	17.77

6. Validation of the pruning procedure on TIMIT and NTIMIT (6s training/3s test - 567 speakers - 2639 tests)

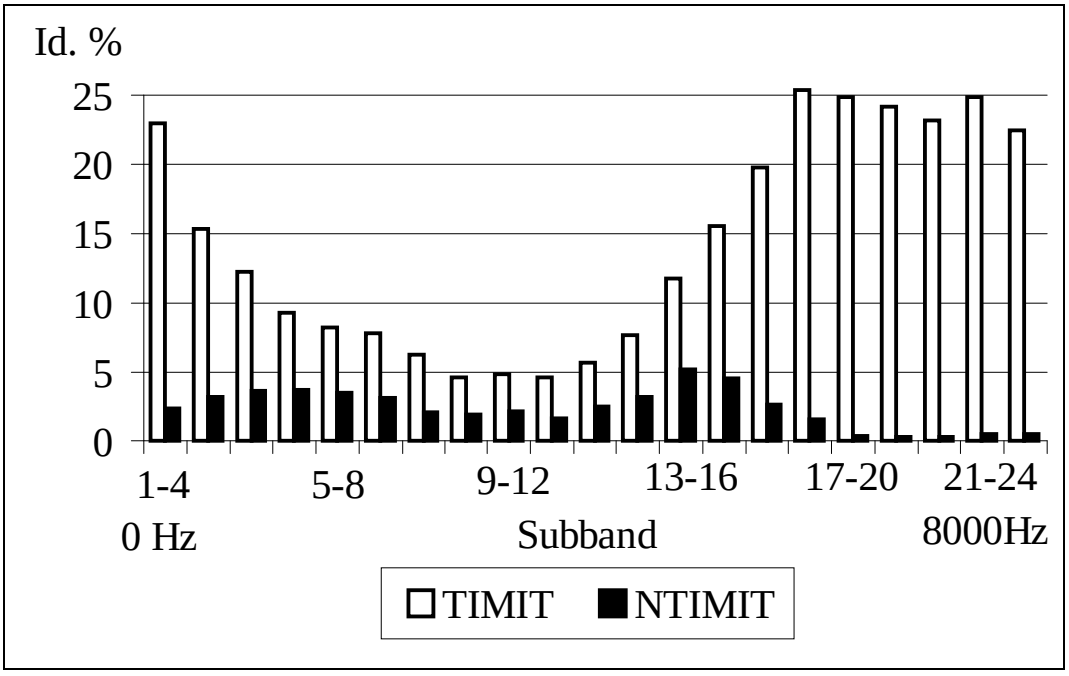
Number of corrupted channels	SNR (dB)	BASELINE n=1; K=1	T-F PRUNING pq=4500; T=1
3	10	13.28	71.67
2	10	23.07	84.26
3	20	44.4	95.1
2	20	58.04	98.25

7. Speaker identification results in the case of speech corrupted by a noise randomly distributed on the whole spectral domain (63 speakers, 286 tests).

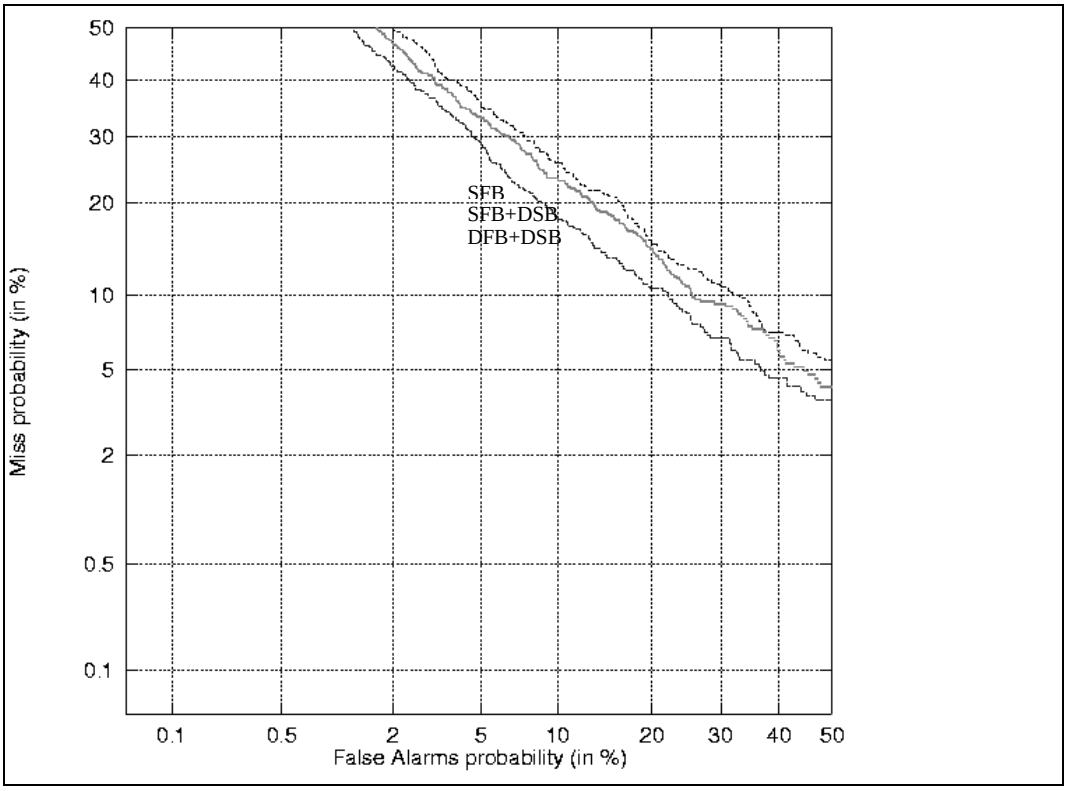
List of Figures.



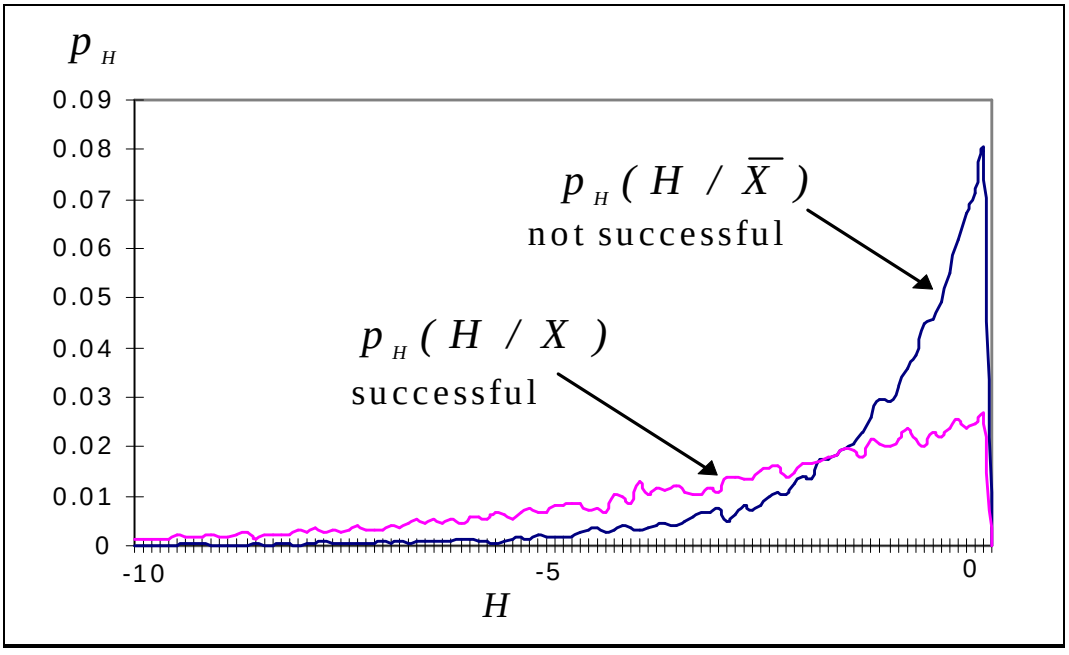
1. speaker identification performance on speech corrupted with noise at different SNRs – MGMs - 24 filterbank coeff. - 6s Training/3s Test – c = number of corrupted channels - [5]



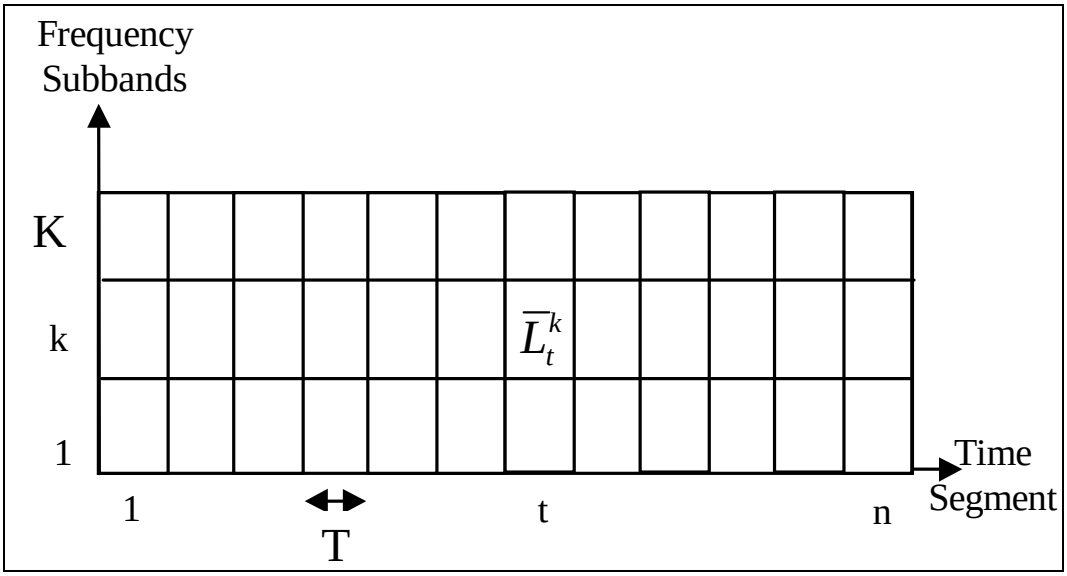
2. isolated subband identification rates on TIMIT and NTIMIT (6s training/3s test - 630 speakers - 2925 tests)



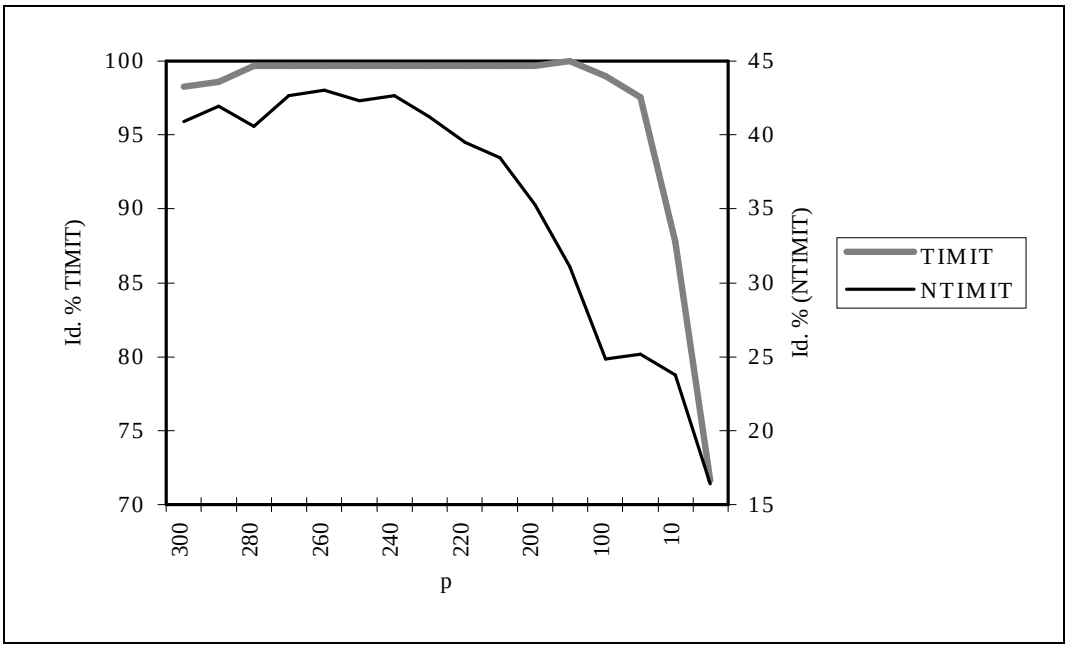
3: DET curves of different recognizer architectures: Static Full Band (SFB), Static Full Band+Dynamic Subbands (SFB+DSB), Dynamic Full Band+Dynamic Subbands (DFB+DSB).



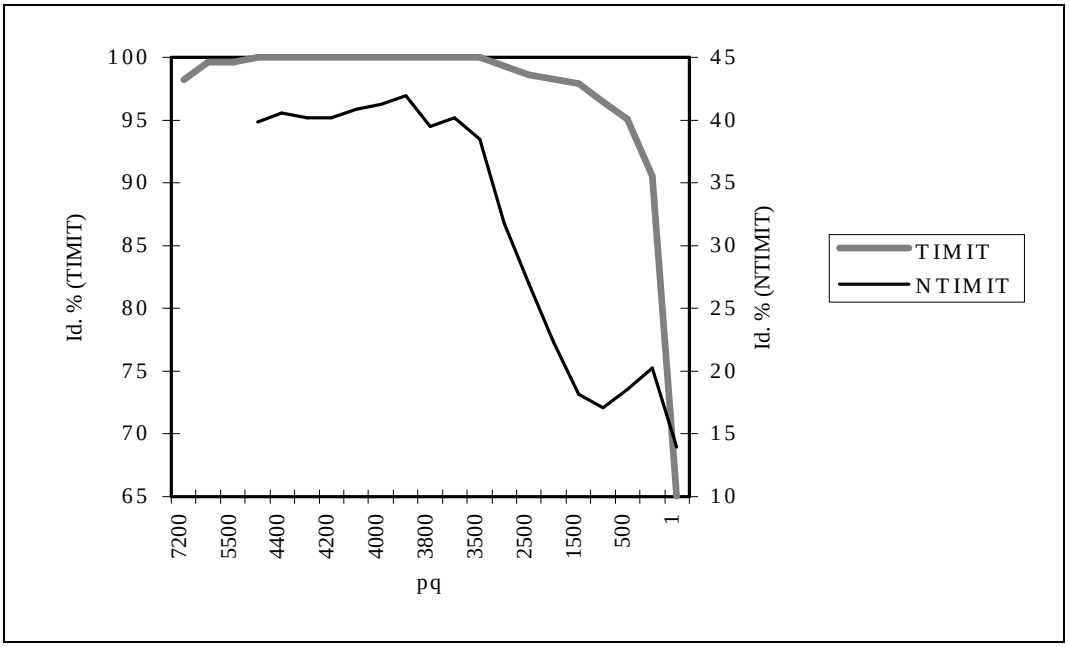
4. Density functions of H_t on the interval [-10,0].



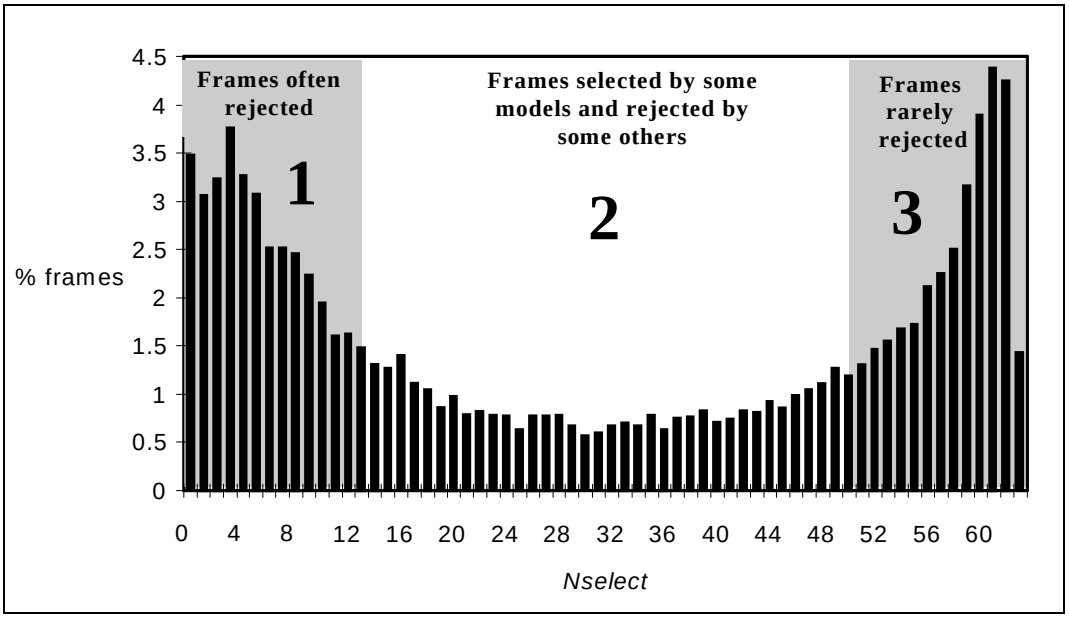
*5. Division of a test utterance into n segments of K subbands (n*K blocks in total)*



6. Time pruning - 6s training/3s test - (300-p) frames pruned - T=1 - 63 speakers - 286 tests



7. Time-frequency pruning - 6s training/3s test - architecture 24x20 for TIMIT and architecture 15x11 for NTIMIT - T=1 - 63 speakers - 286 tests



8. Frame distribution according to their frequency of selection - TIMIT - p=150 (half of the frames rejected) - 63 speakers

[1] Central frequencies of filters (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.

[2] The likelihood of a frame t, uttered by speaker Y is computed from vector y_t and speaker Y's model.

[3] Channels 1,2,3,4,6,8,9,13,15,16,17,18,19,20,21,22,23,24

[4] Channels 1,3,6,13,16,17,18,19,21,22,23,24

[5] Selected features per subband (notation: f-c with f and c referring respectively to frame ($f \in [1,10]$) and channel ($c \in [1,4]$). SB1-4: 1-1, 1-2, 1-3, 1-4, 3-2, 3-3, 3-4, 4-1, 6-1, 9-4. SB5-8: 1-1, 1-2, 1-3, 1-4, 2-3, 2-4, 8-2, 8-3, 8-4, 9-1. SB9-12: 1-1, 1-2, 1-3, 1-4, 2-3, 2-4, 3-2, 3-3, 5-1, 6-1. SB13-16: 1-1, 1-2, 1-3, 1-4, 3-2, 3-3, 4-4, 6-3, 7-2, 9-2, 9-4. SB17-20: 1-1, 1-2, 1-3, 1-4, 2-1, 2-2, 2-3, 2-4, 3-1, 3-4, 4-2, 4-4, 5-2, 6-3, 8-1. SB21-24: 1-1, 1-2, 1-3, 1-4, 2-1, 2-2, 2-3, 2-4, 3-1, 3-3, 3-4, 4-1, 4-4, 5-1, 5-3, 7-3.