

# Time and Frequency Pruning for Speaker Identification

L. Besacier, J.F. Bonastre

LIA/CERI

339, chemin des Meinajaries BP 1228 - 84911 Avignon Cedex 9 (France)

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

## Abstract

This work is an attempt to refine decisions in speaker identification. A test utterance is divided into multiple time-frequency blocks on which a normalized likelihood score is calculated. Instead of averaging the block-likelihoods along the whole test utterance, some of them are rejected (pruning) and the final score is computed with a limited number of time-frequency blocks. The results obtained in the special case of time pruning lead the authors to experiment a joint time and frequency pruning approach. The optimal percentage of blocks pruned is learned on a tuning data set with the minimum identification error criterion. Validation of the time-frequency pruning process on 567 speakers leads to a significant error rate reduction (up to 41% reduction on TIMIT) for short training and test duration.

## 1. Introduction

Mono-gaussian models for speaker recognition have been largely replaced by Gaussian Mixture Models (GMM) which are dedicated to modeling smaller clusters of speech. The Gaussian mixture modeling can be seen as a *cooperation* of models since the gaussian mixture density is a weighted linear combination of uni-modal gaussian densities. The work presented here is rather concerned with *competition* of models since different mono-gaussian models (corresponding to different frequency subbands) are applied to the test signal and the decision is made with the best or the N-best model scores.

More precisely, a test utterance is divided into time-frequency blocks, each of them corresponding to a particular frequency subband and a particular time segment. During the recognition phase, the block scores are accumulated over the whole test utterance to compute a global score and take a final decision. In this work, we investigate accumulation using a hard threshold approach since some block scores are eliminated (pruning) and the final decision is taken with a subset of these scores.

This approach should be robust in the case of a time-frequency localized noise since the least reliable time-

frequency blocks can be removed. Even in the case of clean speech, some speaker test utterance blocks can be simply more similar to another speaker model than to the target speaker model itself. Removing these error-prone blocks should lead to a more robust decision.

In *Section 2*, a formalism is proposed to describe our block-based speaker recognition system. The potential of this approach is shown with a special case of the formalism: time pruning (*Section 3*). Experiments intended to find the optimal percentage of blocks pruned are described in *Section 4*. The optimal parameters (percentage of blocks pruned) are validated on TIMIT and NTIMIT databases (*Section 5*). Finally, we summarize our main results and outline the potential advantages of the time-frequency pruning procedure in *Section 6*.

## 2. Formalism

### 2.1 Mono-gaussian ‘segmental’ modeling

Let  $\{x_i\}_{1 \leq i \leq M}$  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:

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

Similarly, for a speech signal uttered by speaker Y, a sequence of N vectors  $\{y_i\}_{1 \leq i \leq N}$  can be extracted.

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  $y_i$  uttered by speaker Y is:

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

If we assume that all vectors  $y_i$  are independent observations, the average log-likelihood of  $\{y_i\}_{i+1 \leq i \leq i+T}$  on a segment of T frames can be written:

$$\overline{G_X}(y_{i+1}^{i+T}) = \frac{1}{T} \log G(y_{i+1} \dots y_{i+T} / X) = \frac{1}{T} \sum_{i=1}^T \log G(y_{i+1} / X) \quad (3)$$

## 2.2 Multiband modeling

The following ‘K-subbands’ model of speaker  $X$  can be obtained from the initial full-band model:

$$M_X(K) = \left\{ (X^1, \bar{x}^1), \dots, (X^k, \bar{x}^k), \dots, (X^K, \bar{x}^K) \right\} \quad (4)$$

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  $\bar{x}$ . Therefore, the quantities defined in (2) and (3) can be respectively written for the k-th subband:

- $G^k(y_i / X)$  likelihood of vector  $y_i$  on the k-th subband,
- $\overline{G_X^k}(y_{i+1}^{t+T})$  average log-likelihood of segment  $\{y_i\}_{t+1 \leq i \leq t+T}$  on the k-th subband.

## 2.3 Block-based system

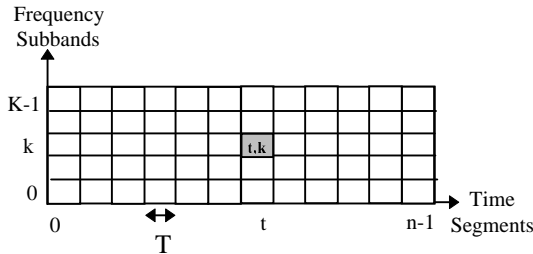
The final model combines both *segmental* and *multiband* aspects. A test utterance of  $N$  frames is divided into  $K$  subbands and  $n$  segments of  $T$  frames each so that  $N=nT$  (Fig. 1). For each pair (t,k) corresponding to the k-th subband of the t-th segment, a normalized score  $h_X^k(y_{i+1}^{t+T})$  (equivalent to a minus log-likelihood ratio) is calculated. This score  $h$  is also called discriminant function [2] (p.52) :

$$h_X^k(y_{i+1}^{t+T}) = \max_{Z \neq X} \overline{G_Z^k}(y_{i+1}^{t+T}) - \overline{G_X^k}(y_{i+1}^{t+T}) \quad (5)$$

The (n\*K) normalized scores are then accumulated over the whole test utterance, to form a final score for each speaker model:

$$\tilde{h}_X(y_1^N) = ACC \left[ h_X^k(y_{i+1}^{t+T}) \right]_{\substack{t \in [0, n-1] \\ k \in [0, K-1]}} \quad (6)$$

where  $ACC$  is the accumulation function ; note that  $h_X(y_1^N)$  is equivalent to the standard gaussian model scoring when  $n=1$  and  $K=1$  (i.e. test utterance considered globally).



**Figure 1: Division of a test utterance in  $n$  segments of  $K$  subbands ( $n*K$  blocks in total)**

The use of different time-frequency blocks enables us to discard or de-emphasize blocks corresponding to abnormal events or blocks poorly representative of the

target speaker. The accumulation function proposed to take advantage of this segmentation is given in (7):

$$\tilde{h}_X(y_1^N) = \arg \min_{(p,q)} \left[ \frac{1}{pq} \sum_1^p \sum_1^q h_X^k(y_{i+1}^{t+T}) \right]_{\substack{t \in [0, n-1] \\ k \in [0, K-1]}} \quad (7)$$

In this case, we average the  $p*q$  lowest block scores for each speaker, with  $p < n$  ( $n$  number of segments in the test utterance) and  $q < K$  ( $K$  number of subbands in the architecture).

This pruning procedure is based on the assumption that the maximum likelihood scores resulting in correct identification are in general higher than the maximum likelihood scores resulting in incorrect identifications. In other words, when a block is error-prone, it is not due to a non-target speaker model matching the speech block well, but rather to the true speaker model performing badly.

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

- if  $K=1$  and  $n > 1$ , we have a "segment level normalization approach" [3] and only time pruning is considered,
- if  $n=1$  and  $K > 1$ , we have a "multiband approach" [1] and only frequency pruning is considered.

## 3. Time pruning

### 3.1 Experimental conditions

For our experiments, we have used TIMIT (normal speech) and NTIMIT (telephone speech) databases. Our speech analysis module extracts filterbank coefficients. The analysis conditions are identical to those used in [1]. For TIMIT database, all 24 coefficients of the spectral vectors are kept. For NTIMIT, we remove the first 2 coefficients and the last 7 coefficients which are outside the telephone band (approximately 300-3400 Hz).

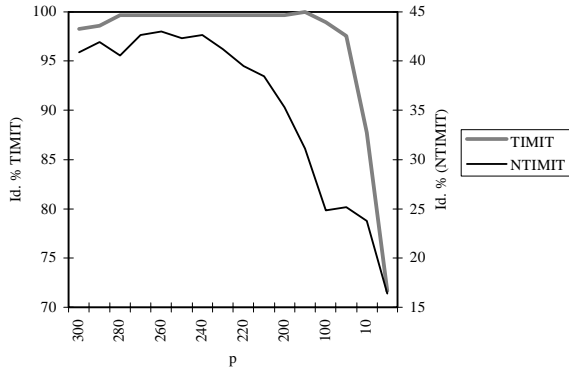
A common training/test protocol is used for all the experiments [1]. Short durations are used (6s training and 3s test) in order to show the efficiency of the pruning procedure even when little speech material is available. All the tests are made within the framework of text-independent closed-set speaker identification using a maximum likelihood decision rule.

### 3.2 Frame pruning

We consider here the special case where  $K=1$  (full-band model); therefore only time pruning is considered. The influence of the number of segments kept  $p$  is investigated when a segment is composed of a single frame ( $T=1$ ). The results are reported in Fig. 2.

For both databases, optimum results are obtained when some frames are pruned: id.=100% for  $p=150$  on

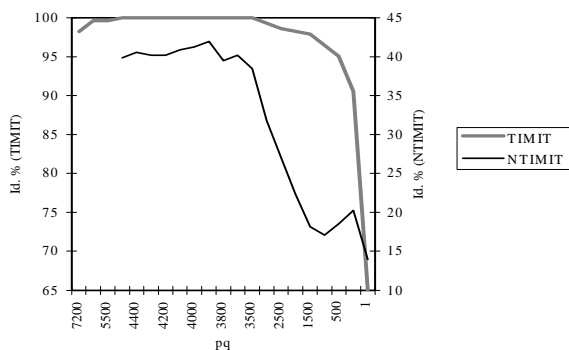
TIMIT and id.=43% for  $p=260$  on NTIMIT. This shows that the selection of the information is important since some frames in a test utterance can contaminate the final score. Moreover, it is interesting to note that a reasonably good performance is obtained on TIMIT when a single frame per speaker is kept (71.63% id.), i.e. when an extremely small amount of speech is used for each speaker to take the final decision !



**Figure 2. Time pruning - 6s training/3s test - (300-p) frames pruned - T=1 - 63 speakers**

#### 4. Time-frequency pruning

We have experimented an architecture of 24 subbands of 20 channels each (24x20) for TIMIT and an architecture of 15 subbands of 11 channels each (15x11) for NTIMIT. The other experimental conditions are the same as those described in Section 3.1. 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 Fig. 3.



**Figure 3. Time and frequency pruning - 6s training/3s test - (7200-pq) or (4500-pq) blocks pruned - T=1 - 63 speakers**

For both databases, 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.

#### 5. Validation

The best values of  $p$  (time pruning, Section 3.2) and  $pq$  (time-frequency pruning, Section 4) obtained for 63 speakers on TIMIT and NTIMIT have been used to validate the benefit of the pruning procedure for speaker recognition. Speaker identification tests have been conducted on the 567 remaining speakers of TIMIT and NTIMIT. So 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 Tab. 1. For both databases, the improvement of performances is significant. The time-frequency pruning procedure leads to a 41% error rate reduction on TIMIT and to 3% error reduction on NTIMIT.

	BASELINE	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

**Table 1. Validation of the pruning procedure on TIMIT and NTIMIT (6s training/3s test - 567 speakers - 2639 tests)**

#### 6. Conclusion

We have presented a time and frequency pruning procedure for speaker identification. The results obtained have shown that this technique can significantly increase the performances of a speaker identification system. We also intend to apply this method to refine the training of the speaker models.

#### 7. References

- [1] BESACIER, L., BONASTRE, J.F., Subband approach for automatic speaker recognition: optimal division of the frequency domain. In *Audio- and Video-based Biometric Person Authentication*, Bigün, et. al. Eds., Springer LNCS 1206, 1997.
- [2] FUKUNAGA, K., *Statistical Pattern Recognition*. Second Edition, Academic Press, Inc., San Diego. 1990.
- [3] MARKOV, K., NAKAGAWA, S., Frame level likelihood normalization for text-independent speaker identification using GMMs. In *Proc. ICSLP*, pp 1764-1767, 1996.