

Independent Processing and Recombination of Partial Frequency Bands for Automatic Speaker Recognition

Laurent BESACIER¹, Jean - François BONASTRE¹

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

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

Abstract. We present the basis of our subband-based speaker recognition approach and preliminary experimental results. 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 selection of the most critical subbands and the recombination strategies are particularly discussed. For the first time, speaker recognition experiments on independent subbands are conducted for 630 speakers, on TIMIT and NTIMIT. The results show that the speaker specific information is not equally distributed among subbands. Speaker identification results similar (even 0.6% better) to the results of a full-band recognition procedure have been obtained on both databases when some subbands were removed from the full-band domain. Even when basic recombination strategies are used, our parallel model slightly outperforms the conventional full-band gaussian measure.

1. Introduction

This article presents a new approach towards automatic speaker recognition. The general principle is to split the whole frequency band into a few sub-bands on which statistical recognizers are independently applied and then recombined to yield global scores and a global recognition decision.

The work described below presents an attempt to (1) describe our subband approach and propose some subband recombination strategies, (2) determine the most critical subbands and select the best frequency channels for speaker recognition, (3) test the systems on TIMIT and NTIMIT databases.

2. Subband Model

2.1 Principle

Our 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.

Subband models for speech recognition have been proposed in [4] and [12]. Some issues involved in designing a subband model can be generalized to speaker recognition :

- the architecture of the subband-based system (selection of the most critical subbands for the recognition task ; optimal division of the whole frequency domain), this part is particularly discussed in a previous article [2].
- the recombination of each subband recognizer output (recombination level, recombination strategies, fusion of multiple decisions).

2.2 Similarity Measure

This measure is more precisely described in [3].

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

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

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

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

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

$$\overline{G_X}(y_1^N) = \frac{1}{N} \log G(y_1 \dots y_N | X) = \frac{1}{N} \sum_{i=1}^N \log G(y_i | X) \quad (3)$$

We define the similarity measure $\mu(X, y_i)$ between a vector y_i uttered by Y and the model of speaker X in order to have :

$$\underset{X}{\text{Arg max}} G(y_i / X) = \underset{X}{\text{Arg min}} \mu(X, y_i) \quad (4)$$

so we define

$$\mu(X, y_i) = -\log G(y_i / X) \quad (5)$$

and similarly for $\{y_i\}_{1 \leq i \leq N}$:

$$\mu(X, y_1^N) = \frac{1}{N} \sum_{i=1}^N \mu(X, y_i) = -\overline{G_X}(y_1^N) \quad (6)$$

This measure is well-adapted to our subband approach : the reference model X^k of speaker X on subband $n^{\circ}k$ is made up of a sub-block of its covariance matrix calculated on the whole spectral domain and of a sub-vector of its mean vector computed on the whole spectral domain.

Finally, for each frame y_i of a test utterance we calculate K measures $\mu^k(X^k, y_i^k)$ each corresponding to the output of the subband recognizer $n^{\circ}k$ ($1 < k < K$).

2.3 Recombination Strategies

2.3.1 Recombination level

The recombination task can occur at different levels [5]:
- 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 [6].

More precisely, the information level at the output of each recognizer can be of three types [6].

(1) ‘distance’ level : a likelihood measure $\mu^k(X^k, y_i^k)$ is computed at the output of each k-th subband recognizer.

(2) ‘sorting’ level : an identification rank $\rho^k(X^k, y_i^k)$ is calculated at the output of each k-th subband recognizer. $\rho^k(X^k, y_i^k)$ is the position where the reference model X^k of speaker X (on subband $n^{\circ}k$) appears in the ordered list of neighbours of test utterance y_i^k .

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

2.3.2 Weighting

Weighting consists in associating a confidence value to each subband recognizer. We have experimented a linear merging technique which corresponds to a weighted sum of the likelihood measures. Three types of weighting have been examined to recombine the scores $s^k(X^k, y_i^k)$ with $s = \{\mu, \rho, \delta\}$ (scores defined in section 2.3.1) :

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

$$s_1(X, y_i) = \frac{1}{K} \sum_{k=1}^K s^k(X^k, y_i^k) \quad (7)$$

ii) weighting according to the accuracy of individual subbands (weights derived from the performance of the individual subbands on a cross-validation data set) :

$$s_2(X, y_i) = \frac{1}{K} \sum_{k=1}^K w_k \cdot s^k(X^k, y_i^k) \quad (8)$$

iii) ‘all or nothing’ weighting ; in this case, we have investigated merging using a hard threshold approach i.e. removing the least reliable subbands (some weights fixed to 0, some others to 1) :

$$s_3(X, y_i) = \arg \min_p \left[\frac{1}{p} \sum_{k \in \{1..K\}} s^k(X^k, y_i^k) \right] \quad (9)$$

2.3.3 Recombination window

Another aspect of our 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 $\{y_i\}_{1 \leq i \leq N}$ of speaker Y .

Actually, the merging task can occur for different time window sizes : recombination after each frame, recombination after x frames, only one recombination at the end of the whole test utterance.

We have only considered the smallest size of window possible, i.e. one recombination procedure after each 10ms frame.

3. Independent processing of partial frequency bands

3.1 Database and signal analysis

For our experiments, we have used TIMIT and NTIMIT databases. TIMIT [6] 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). The NTIMIT database [7] 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 different telephone line for each sentence (local or long distance network). The signal is sampled at 16 kHz, but its useful bandwidth is limited to telephone bandwidth (approximately 300-3400 Hz).

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. Covariance matrices and mean vectors are finally computed from these spectral vectors. These analysis conditions are identical to those used in [2] and [3].

3.2 Training and test protocols

In our protocol, training or test durations are rigorously the same for each speaker.

For the training of a given speaker, all 5 'sx' sentences are concatenated together and the first M samples corresponding to the training duration required (6s here) are selected. Consequently, a single reference pattern is computed from exactly the same number of samples for each speaker.

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. Consequently, the test patterns are computed from exactly the same number of samples for each speaker.

The silences at the beginning and the end of sentences are not removed. All the tests are made within the framework of text-independent closed-set speaker identification using a 1-nearest neighbour decision rule.

3.3 Experiments on isolated subbands

Speaker recognition experiments 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 (6)* applied to each subband.

Fig 1 shows the results obtained for **6s training/3s test on TIMIT and NTIMIT for 630 speakers.**

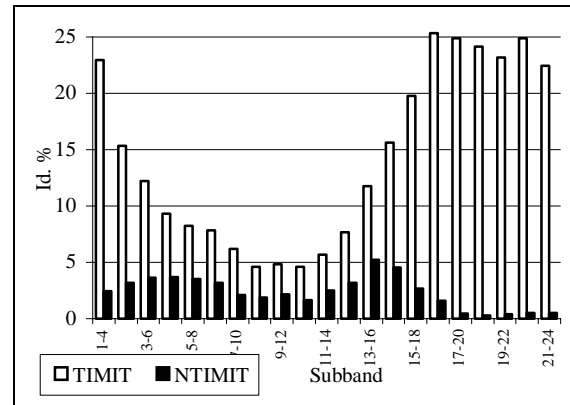


fig 1. Isolated subband identification rates on TIMIT and NTIMIT (6s training, 3s test)

The following comments can be made :

- Large differences between subbands are observed (5% to 25% recognition rates on TIMIT).
- Experiments on TIMIT show that the low-frequency subbands ($f < 600\text{Hz}$) and the high-frequency subbands ($f > 3000\text{Hz}$) are more speaker specific than middle-frequency ones.
- Consequently a drastic performance decrease is observed on NTIMIT for which the most critical subbands are removed (channels 1-2-3 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.
- The high identification rates observed for the first channels on TIMIT could be related to the information conveyed by the fundamental frequency which has been shown to highly contribute to the speaker identification task [1] [9]. This information is partly lost for telephone speech (NTIMIT).

3.4 Channel selection

In this section, 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 [11]. The method begins by evaluating the effectiveness of each of the $N=24$ subbands composed of $N-1$ channels. The most effective subband is then determined, and the channel not included in this subband 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 1. shows the speaker identification rates obtained with the best set of channels on TIMIT and NTIMIT (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.

		FULL BAND	BEST RESULTS	HALF OF THE CHANNELS	HALF OF THE FREQ. DOMAIN
Number of channels	TIMIT	24	18	12	9
	NTIMIT	24	12	12	15
% of the full freq. domain	TIMIT	100.0%	83.2%	64.4%	50.0%
	NTIMIT	100.0%	33.9%	33.9%	50.0%
Id%	TIMIT	93.7%	94.3%	89.5%	79.4%
	NTIMIT	11.9%	17.4%	17.4%	17.2%

Table 1. Main results of the channel selection procedure (630 speakers)

-The best identification results are obtained with 18 channels 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 our current database.

The best identification results are obtained with 12 channels on NTIMIT (17.4%) corresponding to 33.9% of the whole frequency domain. The eliminated channels are those located outside the telephone band as well as channels 9 and 10 (1100Hz<f<1500Hz).

-Good performances are still obtained when using only half of the channels : 89.5% identification rate on TIMIT for 12 well chosen channels ; we can say that the main part of the speaker specific information is condensed in about 60% of the total frequency domain.

4. Recombination experiments

Speaker recognition experiments have been conducted using the multiband technique described in section 2. The tests have been made on a 63-speaker subset of TIMIT (20 women and 43 men). The training and test protocols are the same as those described in section 3. In this experiment, the recombination window size is 10 ms (i.e. one recombination after each 10ms frame). Our parallel model is made up of 24 recognizers consisting of 20 channels each for TIMIT (*Figure 2*). In the case of NTIMIT, the parallel model is made up of 15 recognizers consisting of 11 channels each (*Figure 3*). In fact, for NTIMIT, we have discarded the first 2 channels and the last 7 ones since the useful bandwidth is 330-3400Hz for these data. The global measure at the end of the whole test utterance is an arithmetic mean of recombination scores computed on each frame.

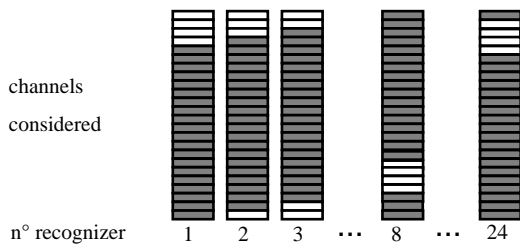


Figure 2 : 24 recognizers of 20 channels for TIMIT

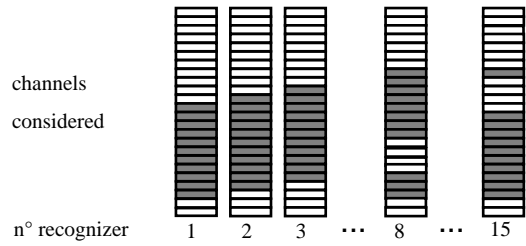


Figure 3 : 15 recognizers of 11 channels for NTIMIT

Table 2 and *Table 3* show the identification results obtained on TIMIT and NTIMIT for different subband recombination strategies.

conventional (full-band)	p-min. (p=1) μ_3	p-min. (p=10) μ_5	equal weighting μ_4	borda ρ_1
97.55	97.9	98.6	97.9	93.35

Table 2. Speaker identification results for different recombination strategies (TIMIT, 63 speakers, 6s training, 3s test, 286 tests).

conventional (full-band)	p-min. (p=1) μ_3	p-min. (p=10) μ_5	equal weighting μ_4	borda ρ_1
40.55	25.87	42.3	37.76	27.27

Table 3. Speaker identification results for different recombination strategies (NTIMIT, 63 speakers, 6s training, 3s test, 286 tests).

In the *p-min* technique, we select only p subbands on which the distance measures between two speakers are the smallest (see μ_3 in section 2). *Borda score* [10] is an alternative voting system based on identification ranks (equivalent to ρ_1). Finally *Conventional* refers to the standard full-band Gaussian likelihood measure (see *equation 6*) computed on the 24 channels for TIMIT and on channels 3 to 17 (useful bandwidth) for NTIMIT.

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. However, even when a basic recombination strategy is used (*equal weighting*), the identification results are similar (even slightly better on TIMIT) to the results of the full-band gaussian measure.

5. Conclusion

In this paper we present the basis of our subband-based speaker recognition approach and preliminary experimental results. The selection of the most critical subbands and the recombination strategies are particularly discussed whereas the choice of an optimal division of the frequency domain (number and size of subbands) is more precisely dealt with in a previous article [1]. For the first time, speaker recognition

experiments on independent subbands are conducted for 630 speakers, on TIMIT and NTIMIT. The results show that the speaker specific information is not equally distributed among subbands. Speaker identification results similar (even slightly better) to the results of a full-band recognition procedure have been obtained on TIMIT and NTIMIT when some subbands were removed from the full-band domain, or when some basic subband recombination strategies were used.

Since we assume that some subbands should perform better for certain classes of speakers than for others, speaker models using speaker-dependent recombination strategies are an interesting issue. These speaker models can 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. Finally, the subband approach can 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 could depend on the phoneme considered.

6. References

- [1] **ATAL, B.S.**, Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America*, n°52, pp 1687-1697, 1972.
- [2] **BESACIER, L., BONASTRE, J.F.**, Subband approach for automatic speaker recognition : optimal division of the frequency domain. In *1st International Conference on Audio and Video Based Biometric Person Authentication (AVBPA)*, March 1997. Crans-Montana (Switzerland).
- [3] **BIMBOT, F., MAGRIN-CHAGNOLLEAU, Y., MATHAN, L.**, Second-order statistical methods for text-independent speaker identification. *Speech Communication*, n°17(1-2), August 1995.
- [4] **BOURLARD, H., DUPONT, S.**, Subband-based speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp 1251-1254, Munich, Germany, April 1997.
- [5] **DASARATHY, B.V.**, *Decision fusion*. IEEE Computer Society Press 1994. Los Alamitos, California.
- [6] **FISHER, W., ZUE, V., BERNSTEIN, J., PALLET, D.**, An acoustic-phonetic database. *JASA*, suppl. A, Vol. 81(S92). 1986.
- [7] **JANKOWSKI, C., KALYANSWAMY, S., BASSON, S., SPITZ, J.**, NTIMIT : a phonetically balanced, continuous speech, telephone bandwidth speech database. *Proc. Internat. Conf. Acoust. Speech Signal Process. '90*, New Mexico, USA, April 1990.
- [8] **LOONIS, P., MENARD, M., BONNEFOY, J.P.**, Fusion d'Informations multi-sources : Etude comparative entre une approche connexionniste dirigée et la règle orthogonale de Dempster-Shafer. *Proc. 10^{ème} RFIA*, pp 606-614, Rennes (France), 16-18 Janvier 1996.
- [9] **MATSUI, T., FURUI, S.**, Text-independent speaker recognition using vocal tract and pitch information. In *Proceedings ICSLP 90*, pp 137-140, 1990.
- [10] **MOULIN, H.**, *Axioms of Cooperative Decision Making*. Cambridge University Press, Cambridge-New-York-Port Chester-Melbourne-Sydney 1988.
- [11] **SAMBUR, M.R.**, Selection of acoustic features for speaker identification. In *IEEE Transactions on ASSP*. n°23(2), pp 176-182, April 1975.
- [12] **TIBREWALA, S., HERMANSTKY, H.**, Subband-based recognition of noisy speech. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp 1255-1258, Munich, Germany, April 1997.