

*Lecture Notes in Computer Science (1206), Audio-
and Video-based Biometric Person Authentication,
Springer LNCS, Bigün, et. al., Eds., 1997.*

Subband Approach For Automatic Speaker Recognition :

Optimal Division Of The Frequency Domain.

Laurent Besacier ^{†,‡} and Jean-François Bonastre [†]

[†] Laboratoire d'Informatique d'Avignon - 339, chemin des Meinajaries BP 1228 -
84140 Avignon Cedex 9 (FRANCE)

[‡] Laboratoire Parole et Langage - Université de Provence 29, av. Robert Schuman,
13621 Aix-en-Provence (FRANCE)

e-mail : besacier@univ-avignon.fr , bonastre@univ-avignon.fr

Abstract. This paper presents a new method for automatic speaker recognition. The principle is to split the whole spectral domain into partial frequency subbands on which recognizers are independently applied and then recombined to yield a global recognition decision. In this article, we particularly discuss the selection of the most critical subbands for the speaker recognition task and the choice of an optimal division of the frequency domain.

Speaker recognition experiments are conducted on different subbands for a 630 population on TIMIT using second-order statistical methods. Large differences in identification between subbands are observed. In particular, the low-frequency subbands (under 600Hz) and the high-frequency subbands (over 2000Hz) are more speaker specific than middle-frequency ones. An appropriate selection of the most critical subbands shows that very good performances are still obtained with only half of the frequency domain.

Finally experiments on different subband system architectures show that the correlations between frequency channels are of prime importance for the speaker recognition task. Some of these correlations are lost when the frequency domain is divided into subbands. Consequently efficient recombination procedures need to be investigated to perform enhanced speaker identification results.

1 Introduction

The work described below falls within the scope of a general study concerning a new method towards automatic speaker recognition. The principle behind this method is to split the whole frequency domain into subbands on which statistical recognizers are independently applied and then recombined to yield global scores and a global recognition decision.

The advantages of a subband approach for automatic speech recognition are reported in [3]. Some of these can be generalized to automatic speaker recognition :

-Some subbands may be more speaker specific than others.

-Different recognition strategies might be applied to different subbands.

The efficiency of a subband method significantly depends on two factors :

(1) *The architecture of the subband-based system* : selection of the most critical subbands for the recognition task ; optimal division of the whole frequency domain (number of subbands, size of subbands).

(2) *The recombination of the output of each subband recognizer* : recombination level, recombination strategies, fusion of multiple decisions.

In this article, point one is investigated in detail. Section 2 is dedicated to the independent processing of partial frequency bands. On that occasion a new test protocol on TIMIT (630 speakers) is described and the results show clearly the most critical subbands for speaker recognition. Slightly enhanced speaker identification results on TIMIT are even obtained by simply removing some bad subbands. Section 3 reports experiments on different subband system architectures in order to find an optimal division of the frequency domain. Finally, Section 4 concludes this work and presents some perspectives for further investigations on the subband-based speaker recognition approach.

2 Independent Processing of Partial Frequency Bands

2.1 Speaker identification measure

The speaker identification measure used in all experiments of this work is inspired from second-order statistical tests on covariance matrices, computed on acoustic parameters [1].

Let X and Y denote two covariance matrices of a reference speaker and of a test speaker respectively, corresponding to the covariance of some spectral vectors computed along a sentence. Let \bar{x} and \bar{y} denote the means of the spectral vectors.

Let M and N denote the number of spectral vectors used to estimate the covariance matrices and mean vectors, and p the dimension of the spectral vectors. The mathematical expression of the measure that we used in our experiments is then :

$$\mu_G(X, Y) = \frac{1}{p} \left[\frac{M}{M+N} \text{tr}(YX^{-1}) + \frac{N}{M+N} \text{tr}(XY^{-1}) - \frac{M-N}{M+N} \log \left(\frac{\det Y}{\det X} \right) \right] + \frac{1}{p} \left[(\bar{y} - \bar{x})^T \left[\frac{M}{M+N} X^{-1} + \frac{N}{M+N} Y^{-1} \right] (\bar{y} - \bar{x}) \right] - 1$$

where 'tr' denotes the trace and 'det' the determinant of a matrix.

Statistical tests on covariance matrices are well adapted to a subband approach. A subband is represented by a subset of q coefficients extracted from the p-dimensional full-band vectors (these coefficients do not need to be consecutive) and each mean vector of a subband is then q-dimensional, while covariance matrices of a subband are q*q symmetric matrices. The advantage is that the covariance matrices of the desired subband can be directly obtained by simply extracting q*q sub-block matrices of the full-band p*p matrices computed once and for all.

2.2 Database and Signal analysis

For our experiments, we used TIMIT database [4] which contains 630 speakers. The speech analysis module extracts filterbank coefficients in the following way : a Winograd Fourier Transform is computed on Hamming windowed signal frames of 31.5ms (i.e. 504 samples) at a frame rate of 10ms (160 samples). For each frame, spectral vectors of 24 Mel-Scale Triangular-Filter Bank coefficients 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 [1].

2.3 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 kept. Consequently, a single reference pattern is computed from exactly the same number of samples for each speaker. The silences at the beginning and the end of sentences are not removed.

For the test of a given speaker, all 'sa' and 'si' sentences (5 in total) are randomly concatenated together and blocks of N samples corresponding to the test duration required are extracted until there is not enough speech data available (limited to a maximum number of blocks per speaker). So the test patterns are computed from exactly the same number of samples for each speaker. All the tests are made within the framework of text-independent closed-set speaker identification using a 1-nearest neighbour decision rule.

2.4 Experiments on isolated subbands

Speaker recognition tests have been conducted on 21 subbands consisting of 4 consecutive channels (4-dimensional subbands) with band-overlap (subband 1 : channels 1 to 4, subband 2 : channels 2 to 5, ... , subband 21 : channels 21 to 24).

Fig. 1 shows the results obtained for 3s test and 6s test on TIMIT database (6s training - 630 speakers). In order to compare the relative performances of subbands to phonetic events, the identification rates and the total of the first 3 formant value distributions are superposed on the same frequency scale. The vowel formant values are taken from the literature [5] concerning 5 different experiments.

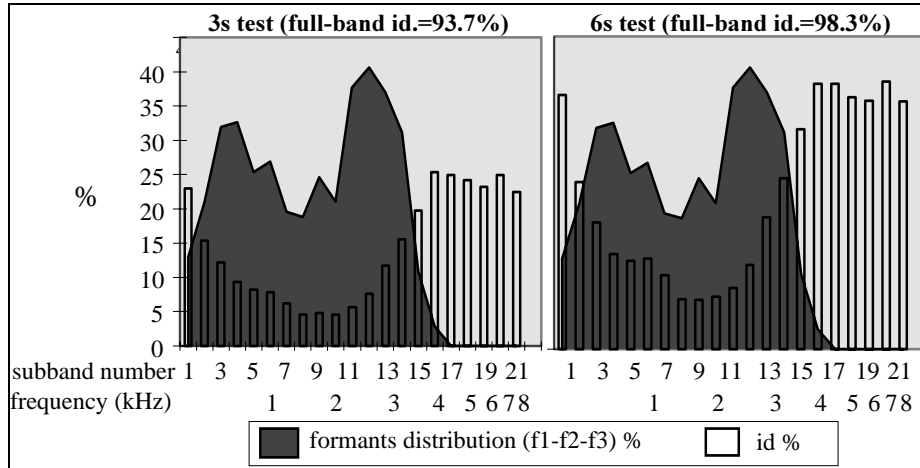


Fig. 1 Subband identification rates (TIMIT-630) compared to formant value distribution

-Large differences between subbands are observed (7% to 38% recognition rates for 6s test on TIMIT).

-The speaker recognition rates on individual 4-dimensional subbands can reach 25% for 3s test which is rather high for a single subband and a 630 population database.

-The low-frequency subbands ($f < 600\text{Hz}$) and the high-frequency subbands ($f > 2000\text{Hz}$) are more speaker specific than middle-frequency ones (whatever the time duration). These results explain the drastic performance decrease observed for telephone quality [8] since the most critical subbands are removed in this case.

-The lowest identification rates correspond to a less important formant distribution (between 1st and 2d formant), which reinforce the idea that formants convey intrinsic speaker characteristics [7][2]. However, it can be seen that speaker specific information is also present beyond the third formant. This higher frequency information may be conveyed by fricatives and higher-order formants.

-The high identification rates observed for the first channel are related to the information conveyed by the fundamental frequency which has been shown to highly contribute to the speaker identification task [6].

2.5 Removing some subbands from the full-band frequency domain

In the previous experiments, speaker-specific information was shown to be present in all subbands even if large differences between them were observed. This section is an attempt to categorize these different subbands in the sense of a subband-based speaker recognition approach. For this, speaker recognition tests have been conducted (TIMIT630 - 3s test) on the full-band frequency domain from which a subband was systematically discarded.

Table 1 shows identification results obtained for each 20-dimensional subband (where a 4-dimensional subband is discarded). Recombination results of each 20-dimensional subband with its dual 4-dimensional subband are also reported.

The recombination score is basic : it corresponds to the arithmetic mean of the distance measures computed on both subbands.

Three classes of subbands can be observed :

-Some results obtained on 20-dimensional subbands are better than those obtained on the full-band frequency domain (results > 93.7% ; bold figures in the 2d column). In this case, the term *disruptive subband* could be used to qualify the subbands discarded since the identification rates are better without these subbands. We note that the disruptive subbands correspond to the subbands which independently lead to the lowest identification rates (Section 2.4).

-The term *average subband* could be used to qualify non-disruptive subbands which do not improve the identification rates either when combined with their dual 20-dimensional subband.

-Finally, the term *critical subband* could be used to qualify a subband which improves the identification rates when combined with its dual band (recombination results > 20-dimensional individual subband results ; bold figures in the 3rd column).

channels discarded	id% (20-dimensionnal subband)	recomb. id%
1 to 4	89.0	90.8
2 to 5	90.8	88.4
3 to 6	91.6	87.8
4 to 7	92.5	87.6
5 to 8	93.0	89.5
6 to 9	93.4	88.3
7 to 10	93.5	86.9
8 to 11	93.8	86.0
9 to 12	94.1	86.2
10 to 13	94.0	86.5
11 to 14	93.9	85.8
12 to 15	93.0	85.9
13 to 16	91.7	86.9
14 to 17	88.8	87.5
15 to 18	87.2	88.1
16 to 19	86.2	87.4
17 to 20	84.8	85.3
18 to 21	83.6	82.7
19 to 22	83.2	83.0
20 to 23	82.5	83.7
21 to 24	82.3	85.4

Table 1. Identification results obtained for 20-dimensional individual subbands and for each recombination of a 20-dimensional subband with its dual 4-dimensional subband

Number of channels	% of the full freq. domain	id %	mean of intra speaker dist.	mean of inter speaker dist.	channel removed
24	100	93.7	0.52	1.63	10
23	97.3	94.0	0.52	1.66	5
22	95.6	94.1	0.50	1.65	11
21	92.5	94.2	0.49	1.66	12
20	89.3	94.2	0.47	1.66	14
19	85.3	94.3	0.45	1.66	7
18	83.2	94.3	0.44	1.69	9
17	80.7	94.0	0.43	1.71	15
16	76.4	93.4	0.40	1.71	20
15	69.6	92.8	0.38	1.70	8
14	67.3	91.9	0.37	1.72	2
13	66	91.1	0.36	1.55	4
12	64.4	89.5	0.34	1.56	13
11	60.8	86.8	0.31	1.57	6
10	59	83.9	0.29	1.54	23
9	50	79.4	0.27	1.59	3
8	48.5	73.0	0.26	1.26	16
7	43.8	64.8	0.25	1.24	22
6	35.6	55.2	0.22	1.19	19
5	29.4	42.7	0.20	1.17	18
4	23.7	28.5	0.18	1.11	24
3	13.9	16.0	0.16	1.08	17
2	8.7	7.0	0.14	1.08	21
1	1.2	2.5	0.09	1.48	1

Table 2. 24 steps of the knock-out procedure

2.6 'Knock-out' procedure for subband selection

In the previous section, it was shown that slightly enhanced speaker identification results on TIMIT could be obtained by simply removing some disruptive subbands. In this section, a subband selection method is proposed to estimate more precisely the relative effectiveness of each part of the frequency domain.

The method used is the well known 'knock-out' procedure [9]. The method begins by evaluating the effectiveness (identification performance) 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') from further consideration. The procedure continues until all the channels are eliminated from consideration. The ordered effectiveness of the channels is then given by the reversed sequence of 'knocked-out' channels.

Experiments were conducted on TIMIT database (630 speakers - 6s training - 3s test). The 'knocked-out' method required the same test procedure as the one described in Section 2.3 for $N(N+1)/2=300$ different subband systems. *Table 2* shows each step of the knock-out procedure. The percentage of the whole frequency domain occupied by the remaining channels is given. The global means of the intra and inter-speaker distances, the identification results and the channel removed after each step are also reported in this table.

The following comments can be made :

-The best identification results are obtained with 18 channels (94.3% identification) 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% identification). However, this improvement may be only considered as an a-posteriori optimization of the results on our current database.

-As expected, the disruptive channels are the middle-frequency ones (channels 10,5,11,12,14,7) whereas the critical channels are the high frequency ones (16,22,19,18,24,17,21). The best channel is channel one : experiments on this channel alone gave 2.5% identification rate but we speculate about the behavior of second-order statistical measures on a single channel.

-We also note that the best identification results approximately correspond to the biggest means of the inter-speaker distances ; the inter-speaker mean is bigger when the distances between speakers are computed on 18 channels rather than on the 24 channels. In a way, we can feel free to think that the disruptive subbands neutralize the ability of the other subbands to separate speakers.

-As opposed to this, the mean of the intra-speaker distances is always bigger when the number of channels used to compute the measure increases.

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

3 Different Subband System Architectures

The choice of an optimal division of the frequency domain seems to be crucial for any subband approach. In this section, experiments on different subband system architectures are presented. The independent processing and the recombination of partial frequency bands have been conducted for different divisions of the frequency domain : 12 subbands of 2 channels ; 8 subbands of 3 ; 6 subbands of 4 ; 4 subbands of 6 ; 3 subbands of 8 and 2 subbands of 12. Database, signal analysis and training/test protocols are the same as those described in Section 2. The channels in a subband are either consecutive or crisscrossed. *The final recombination score corresponds to the mean of the distance measures computed on each subband.*

	Independent Subbands											Recomb. id%	
channels	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-16	17-18	19-20	21-22	23-24	consecutive
id %	7.6	3.8	2.5	2.3	2.0	1.8	1.8	3.4	4.5	5.3	5.7	4.6	58.7
channels	1-13	2-14	3-15	4-16	5-17	6-18	7-19	8-20	9-21	10-22	11-23	12-24	crisscrossed
id %	5.6	2.5	2.6	2.5	2.2	2.6	2.3	2.8	2.9	2.5	2.9	2.4	24.1
channels	1-2-3	4-5-6	7-8-9	10-11-12	13-14-15	16-17-18	19-20-21	22-23-24					consecutive
id %	16.6	5.3	5.2	3.0	5.1	13.4	13.1	10.8					73.1
channels	1-9-17	2-10-18	3-11-19	4-12-20	5-13-21	6-14-22	7-15-23	8-16-24					crisscrossed
id %	9.2	4.6	5.2	4.5	4.4	5.4	6.5	6.2					29.0
channels	1-2-3-4	5-6-7-8	9-10-11-12	13-14-15-16	17-18-19-20	21-22-23-24					consecutive		
id %	23.0	8.2	4.8	11.8	24.9	22.5					80.5		
channels	1-7-13-19	2-8-14-20	3-9-15-21	4-10-16-22	5-11-17-23	6-12-18-24					crisscrossed		
id %	14.4	10.3	9.7	9.4	11.2	11.4					41.2		
channels	1-2-3-4-5-6	7-8-9-10-11-12	13-14-15-16-17-18	19-20-21-22-23-24					consecutive				
id %	30.9	8.5	33.1	48.0					85.2				
channels	1-5-9-13-17-21	2-6-10-14-18-22	3-7-11-15-19-23	4-8-12-16-20-24					crisscrossed				
id %	30.5	25.7	27.5	29.0					58.4				
channels	1-2-3-4-5-6-7-8	9-10-11-12-13-14-15-16	17-18-19-20-21-22-23-24					consecutive					
id %	39.0	20.5	68.5					88.3					
channels	1-4-7-10-13-16-19-22	2-5-8-11-14-17-20-23	3-6-9-12-15-18-21-24					crisscrossed					
id %	47.9	44.5	46.6					71.6					
channels	1-2-3-4-5-6-7-8-9-10-11-12	13-14-15-16-17-18-19-20-21-22-23-24					consecutive						
id %	40.8	83.7					90.6						
channels	1-3-5-7-9-11-13-15-17-19-21-23	2-4-6-8-10-12-14-16-18-20-22-24					crisscrossed						
id %	75.1	72.0					83.5						

Table 3. Independent subband identification results and recombination results for different divisions of the frequency domain (3s test - TIMIT - 630 speakers - chance threshold=0.16% - full-band id.=93.7%)

Table 3 shows identification results independently obtained for each size of subband. Recombination results obtained for different divisions of the frequency domain are also presented in this table. The following observations can be made :

- The independent results obtained with subbands consisting of consecutive channels confirm the lack of speaker specific information observed for the middle frequencies. This is particularly clear for the 6-dimensional subbands since the identification results are only 8.5% for channels 7 to 12 against 30.9% for channels 1 to 6 and 33.1% for channels 13 to 18.
- The identification rates obtained on subbands having the same number of crisscrossed channels are approximately the same. In this case, the information conveyed by different subbands is redundant because each band is composed of channels regularly distributed on the whole spectral domain.
- The recombination results obtained with consecutive channels are far better than those obtained with crisscrossed channels (73.1% against 29% for 3-dimensional subbands). These results show that the correlations between close channels are important for the speaker recognition task when second-order statistical methods are used. Actually, the correlations between a channel and its close neighbours are discarded when a subband is composed of crisscrossed channels.
- The differences of recombination performances between consecutive and crisscrossed channels logically decrease when the subband size is bigger since there are less correlations lost through crisscrossing in this case.
- We also note that the recombination performances decrease when the subbands become smaller and more numerous (whether the channels are crisscrossed or consecutive). In fact, when the subbands are smaller, less parameters are used to model a speaker (Table 4). The deleted parameters correspond to the correlations between distant channels.

number of subbands (n)	1	2	3	4	6	8	12
size of subbands (p)	24	12	8	6	4	3	2
total number of parameters used (np ²)	576	288	192	144	96	72	48

Table 4. Number of parameters used to model a speaker for different divisions of the frequency domain

4 Conclusion

In this paper, we presented a new speaker recognition approach based on independent processing and recombination of partial frequency subbands. For the first time, speaker recognition experiments on independent subbands have been conducted for 630 speakers on TIMIT database. The results have shown that the speaker specific information is not equally distributed among subbands. These differences of performance between subbands have been related to phonetic events. Speaker identification results similar to the results of a full-band recognition procedure have been obtained when some subbands were removed from the full-band domain. Experiments on different subband system architectures have shown that the correlations between frequency channels are of prime importance for the speaker recognition task. Some of these correlations are lost when the frequency domain is divided into subbands ; thus efficient recombination strategies need to be developed. 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 could be applied to speaker verification. In this case, the verification task could be performed on the optimal bands of the applicant speaker.

The subband approach can also be developed at the signal analysis level and different signal-processing tasks might be applied to different subbands. For instance, the length of the analysis window could be large for the low frequencies and narrow for the high frequencies (different resolutions in order to track different speech events). In this case, the speech-processing would not be far from the wavelet transform for which analyzing objects have different properties according to their location in the time-frequency domain.

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.

Acknowledgments

We are indebted to Frederic Bimbot for his advice and criticism about the work presented here

5 References

- [1] BIMBOT, F., MAGRIN-CHAGNOLLEAU, Y., MATHAN, L., Second-order statistical methods for text-independent speaker identification. *Speech Communication*, n°17(1-2), pp 177-192, August 1995.
- [2] BONASTRE, J.F., MELONI, H., Inter and intra-speaker variability of french phonemes ; advantages of an explicit knowledge based approach. *In Workshop on Automatic Speaker Recognition*, pp 157-160, April 1994. Martigny (Switzerland).

- [3] BOURLARD, H., DUPONT, S., A new ASR approach based on independent processing and recombination of partial frequency bands. *In Proceedings ICSLP*, October 1996. Philadelphia, USA.
- [4] FISHER, W., ZUE, V., BERNSTEIN, J., PALLET, D., An acoustic-phonetic database. *JASA*, suppl. A, Vol. 81(S92). 1986.
- [5] HOLLIEN, H., *The acoustics of crime*. Applied Psycholinguistics and Communication Disorders 1990. Plenum Press : New-York & London. 370p.
- [6] MATSUI, T., FURUI, S., Text-independent speaker recognition using vocal tract and pitch information. *In Proceedings ICSLP 90*, pp 137-140, 1990.
- [7] NOLAN, F., *The phonetic bases of speaker recognition*. CUP 1983. Cambridge.
- [8] REYNOLDS, D.A., Speaker Identification and verification using gaussian mixture models. *In Workshop on Automatic Speaker Recognition and Verification*, pp 27-30, April 1994. Martigny (Switzerland).
- [9] SAMBUR, M.R., Selection of acoustic features for speaker identification. *In IEEE Transactions on ASSP*. n°23(2), pp 176-182, April 1975.