LDA Combination of Pitch and MFCC Features in Speaker Recognition

A. Harrag[1], T. Mohamadi[1] and J. F. Serignat[2]

Abstract – **This paper assesses two popular speaker features pitch and cepstral coefficient (static and dynamic). We compare the performance of individual features and features combined via LDA. The identification process can be performed both in the temporal and cepstral domains. The temporal analysis determines which phonemes or utterances exhibit the highest degree of speaker specificity, while the cepstral analysis examines individual cepstra within these temporal divisions.**

Keywords **– Speaker recognition, Intra-speaker variation, Inter-speaker variation, cepstral coefficient, F-ratio, linear discriminant analysis.**

I. Introduction

Investigations into the phonetic aspects of speech in speaker recognition are reported by [1-2]. In those investigations, it has been found that speaker specificity varies noticeably between different phonetic subgroups, but remains more constant within them. These phoneme subgroups can be crudely ranked in the following order: long vowels, nasals, short vowels, fricatives, and plosives, where long vowels exhibit the highest degree of speaker specificity, and the plosives the lowest.

The speech database consists of 10 speakers (6 female and 4 male), uttering the vowels (a, i, e, u and o) four times. The sampling rate of the data is 16 KHz. This corpus is extracted from BDSONS the French speech database.

II. Experimentations

A.Pitch

For this parameter we calculated the average value and the standard deviation on each realization of each vowel for all the 10 speakers. According to the obtained results, we note that this parameter does not allow a good discrimination between speakers, nevertheless it allows a first subdivision of space in two distinct classes: female and male. The results obtained are presented in Table 1, Table 2, Fig. 1 and Fig. 2.

B.Cepstral Coefficients

The cepstral features considered here are generated from speech database using a 10-channel DFT-simulated Mel filter-bank [3], with a 95% overlap between successive frames. This filter bank is widely used and leads to the standard for the Mel cepstra.

Table I: Mean value

	Average value				
	/a/	/e/	/i/	/o/	/u/
Loc1	99,91	103,15	110,56	108,085	109,305
Loc2	223,13	231,22	229,13	233,12	226,34
Loc3	211,71	219,69	222,53	222,05	221,67
Loc4	104,90	111,92	115,24	114,93	115,43
Loc5	228,71	241,95	247,84	251,14	250,17
Loc6	202,89	212,662	247,24	229,38	249,35
Loc7	130,42	134,25	131,96	129,49	132,69
Loc8	207,19	209,98	213,88	209,60	213,53
Loc9	112,42	115,88	126,25	120,14	120,26
Loc10	257,91	262,68	276,01	270,63	281,56

Table II: Standard deviation

	Standard deviation				
	/a/	/e/	/i/	/o/	/u/
Loc1	2,28	3,45	3,37	2,02	2,91
Loc2	6,21	6,06	3,072	3,92	4,83
Loc3	6,39	8,38	7,65	7,43	5,83
Loc4	6,23	6,18	7,41	6,22	6,45
Loc5	9,13	12,57	12,75	15,42	11,91
Loc6	6,99	6,21	5,76	6,88	4,77
Loc7	3,84	4,74	4,71	4,40	3,56
Loc8	3,13	4,83	2,64	5,31	2,97
Loc9	3,08	2,88	2,15	2,37	2,63
Loc10	4,35	4,79	4,95	5,94	7,21

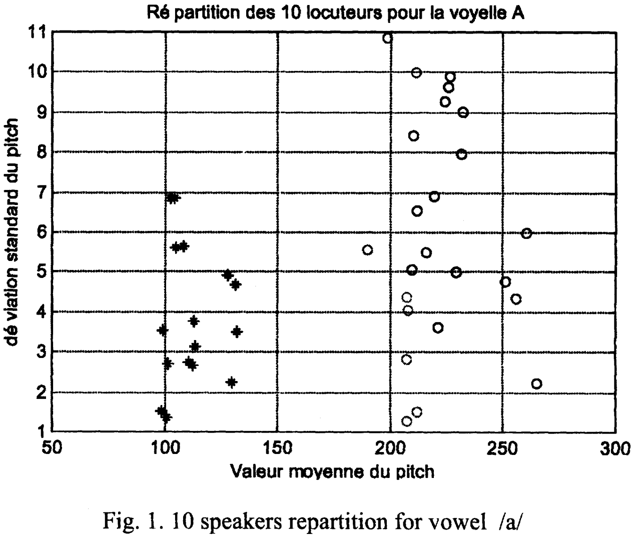

Fig. 1. 10 speakers repartition for vowel /a/

[1] Institut d'Electronique, Universite Ferhat Abbas, CP 19000 Setif Algerie,

[2] Laboratoire CLIPS, Grenoble, France;
Email: aharrag@free.fr Tel : +33 6 20 77 56 00

0-7803-9503-4/05/$20.00 ©2005 IEEE

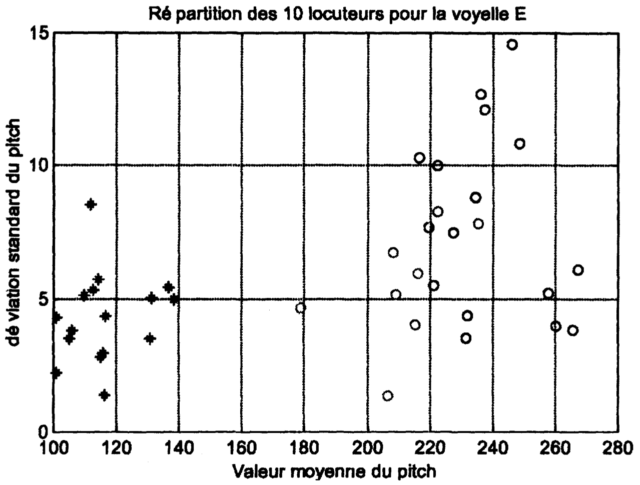

Fig. 2. 10 speaker repartition for vowel /e/

For these parameters, we used like standing method of F-ratio [4], defined by equation (1):

$$F_{ratio} = \frac{\sigma^2_{inter}}{\sigma^2_{intra}} \qquad (1)$$

Where

F_{ratio} : inter to intra – variance ratio

σ^2_{inter} : variance of talker means

σ^2_{intra} : average within talker variance

The obtained results are presented below, for each type of parameters, we present the tables (Tables 3,4 and 5) of the F-ratio and two figures which represent the distribution of the 10 speakers with the corresponding parameter from the highest F-ratio (Fig. 3) to the lowest F-ratio (Fig. 4) for each vowel.

A.Analysis of combined LDA features

Many people have applied LDA in the context of speech recognition where vocabulary items are defined as classes [5]. Here we applied LDA in speaker recognition and a class is defined for each speaker [6-7].

LDA can be summarised as a two-phase procedure: class-dependent normalisation, followed by a discriminative optimization. In phase one, a class-dependent normalisation function collects statistical information via an averaged intra-class covariance matrix W over all speakers.

Table. III: F-ratio for MFCC parameters

	/a/	/e/	/i/	/o/	/u/
C_1	0,9824	2,4206	2,2671	1,3818	1,8293
C_2	1,1443	3,1187	**5,1243**	2,0883	4,6613
C_3	**2,2851**	2,4698	2,2491	2,6773	2,4866
C_4	1,0280	3,0492	2,3223	1,8988	**1,4442**
C_5	1,9000	**3,2963**	3,0096	2,3954	3,2915
C_6	1,1479	**1,9420**	2,9280	**1,1503**	3,5302
C_7	0,7389	3,2243	3,0492	1,2787	**6,2686**
C_8	0,6429	2,8752	3,8096	3,3202	4,2763
C_9	0,8217	3,0588	4,1155	**6,9584**	4,6085
C_{10}	**0,4990**	2,4458	**2,2347**	2,6605	6,0575

Table. IV. F-ratio for DMFCC parameters.

	/a/	/e/	/i/	/o/	/u/
C_1	0,8241	**1,4694**	0,4654	**1,7741**	0,5314
C_2	**6,5373**	0,4126	0,2795	0,3587	0,4906
C_3	1,5078	0,4306	0,3623	0,2927	**0,3994**
C_4	0,7521	0,9320	0,3826	0,4750	0,4930
C_5	0,6273	0,3239	0,3754	0,2471	1,1107
C_6	**0,2555**	0,4150	0,4858	0,5746	0,6645
C_7	0,3934	0,6981	**0,2507**	0,5062	0,9488
C_8	0,3311	0,9440	**0,7269**	0,4474	**1,2211**
C_9	0,6429	**0,2471**	0,3635	0,4282	0,7677
C_{10}	0,3575	0,6645	0,6741	**0,1080**	0,6945

Table.V. F-ratio for ΔΔMFCC parameters

	/a/	/e/	/i/	/o/	/u/
C_1	**0,5710**	**0,7989**	2,6617	1,3111	**1,0184**
C_2	0,9776	**3,2687**	**7,6205**	2,6557	5,9040
C_3	1,3734	2,8129	3,0683	2,9796	2,6905
C_4	1,5906	2,9400	2,8212	1,5594	1,1719
C_5	**3,4642**	2,4482	4,8088	2,5777	3,3226
C_6	2,0524	1,5894	3,8468	1,3602	3,2819
C_7	1,3878	3,5098	3,6525	**1,3099**	**8,5501**
C_8	0,7773	2,5466	3,1763	3,4102	5,9640
C_9	0,8265	2,9772	**2,4326**	**5,3942**	3,6249
C_{10}	0,9452	2,8164	3,4258	3,2951	6,5913

$$W = \frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{j=1}^{N_t}(V_{ij} - M_i)(V_{ij} - M_i)^T \qquad (2)$$

In phase two, a discrimination function is derived from the inter-class covariance matrix B defined by equation (3).

N_s : number of the speakers

N_t : number of the training feature vectors per speaker

V_{ij} : feature vector j for speaker i

M_i : mean of the training feature vectors for speaker i

In phase two, a discrimination function is derived from the inter-class covariance matrix B defined by equation (3).

$$B = \frac{1}{N_s}\sum_{i=1}^{N_s}(M_i - M)(M_i - M)^T \qquad (3)$$

M : global mean of all training data for all speakers

The resultant LDA feature contains elements that are uncorrelated, ranked according to the objective criterion computed from the inter-class covariance matrix B and the averaged intra-class covariance matrix W i.e. $W^{-1}B$.

In term of the principal component analysis, the elements with high ratio contain the most information for speaker recognition, and ones with low ratio have less or even irrelevant information so that the top best elements can be adopted, discarding lower ranked ones to give a reduced feature dimension.

LDA maximizes the intra-class separation and decreases the inter-class dispersion. Contrary, to the selection per F-ratio which evaluates the individual performance of each parameter, the LDA analysis works on vectors of parameters by making the assumption that each new parameter brings more information to the classification process (i.e. without eliminating the coefficients with poor performances).

We used different combinations of some features to show that LDA identifies redundancies and leads to a reduction of the overall dimensions. The combinations are as follows:

- Pitch + 5 MFCC
- Pitch + 10 MFCC
- Pitch + 10 MFCC + 5 ΔMFCC
- Pitch + 10 MFCC + 10 ΔMFCC
- Pitch + 10 MFCC + 10 ΔMFCC + 5 ΔΔMFCC
- Pitch + 10 MFCC + 10 ΔMFCC + 10 ΔΔMFCC

Considering the significant number of the figures (36, six per combination), we just give some examples with utterance /a/ to show the obtained results (Fig. 5).

According to the results and the figures presented below, we note that the assumption made above is verified, i.e. each time a new vector of parameters is added, the inter-class separation is improved and intra-class dispersion is minimized.

III. Conclusions

The pitch parameter is used to separate the speakers in two quite distinct classes (female and male). Unfortunately, this parameter does not allow a clear discrimination between speaker, so we need coupling it with other more discriminating parameters such as the MFCC's and their first and second derivative.

Cepstral coefficients are statistically very slightly correlated. This property makes useless the procedures of orthogonalisation often employed with other sets of parameters; on the other hand, it simplifies the measurement of the distance between two sets of cepstral parameters.

The cepstral parameters are a simple mean to identify the difference in levels, which can exist between two identical spectra, but, we does not take into account the first cepstral coefficient due to level standardization of our data.

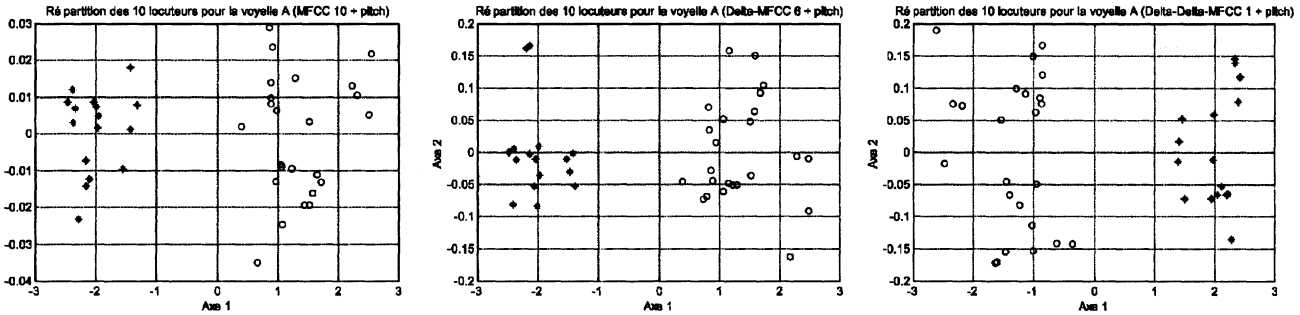

Fig 3. Highest F-ratio for parameters: a) MFCC; b) ΔMFCC et c) ΔΔMFCC.

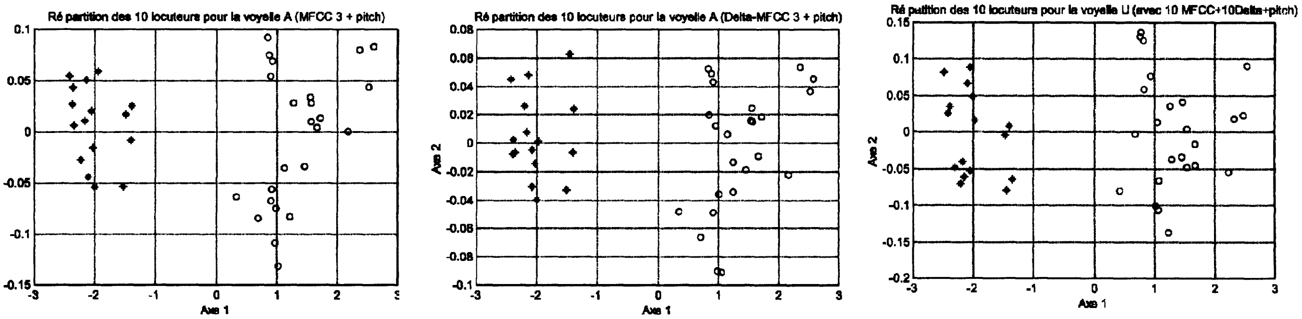

Fig 4. Lowest F-ratio for parameters: a) MFCC; b) ΔMFCC et c) ΔΔMFCC.

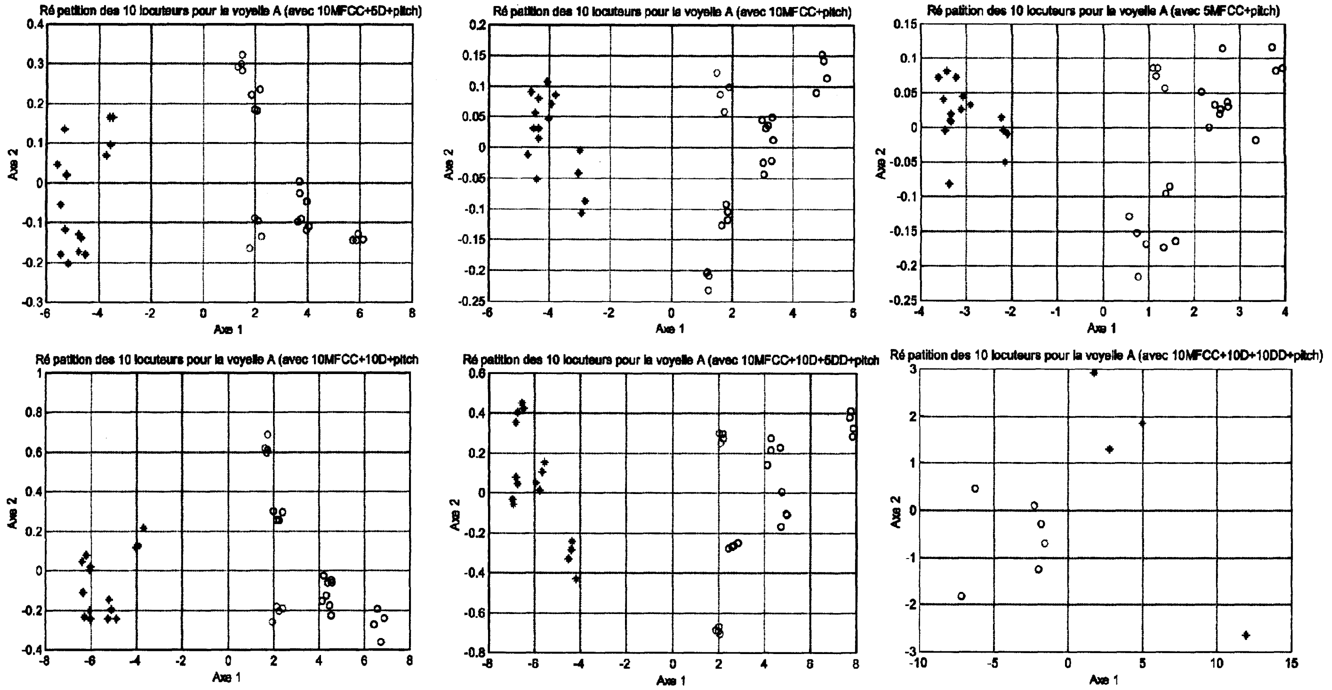

Fig. 5. Speaker repartition for the 6 LDA combinations features.

For LDA, we noted that the linear discriminating analysis makes possible to transform the space of origin towards another space more discriminated, except that this analysis was made on the whole of the vector of entry i.e. without reduction of dimension and elimination of the parameters which degrade the performances of the systems. For that, it is necessary to take into account the results obtained per F-ratio on each parameter and which individually gives an idea to the measure of performance of each parameter.

References

[1] Eatock J., "Speech classification and phoneme performance in speaker recognition", Ph. D. Thesis, University College, Swansea, 1992.

[2] Heuvel Van den and Reitveld R., "Speaker related variability in cepstral representation of Dutch speech segments", ICSLP-92, Canada, Vol 2, pages 1581-1584, 1992.

[3] Furui S., "Digital speech processing, synthesis and recognition", Markel Dekker, New York, 1989.

[4] Pruzansky S. and Mathews M. V.,"Talker recognition system based on the analysis of the variance", J. Acoust. Soc. Am., Vol. 36, No. 11, pp.2041-2047, November 1964.

[5] Doddington. G. R.,"Phonetically sensitive discriminants for improved speech recognition", ICASSP-89, pages 556-559, 1989.

[6] Genoud D., "Reconnaissance et transformation de locuteurs", Thèse de Doctorat, Université de Lausanne, 1999.

[7] Mami Y., "Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence", Thèse Doctorat, ENST Paris, France, 2003.