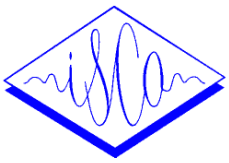

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

AR-VECTOR MODELS FOR FREE-TEXT SPEAKER RECOGNITION

Claude Montacié & Jean-Luc Le Floch

LAFORIA - Université Paris 6, CNRS-URA 1095, 4, place Jussieu, 75252 Paris Cedex 5

ABSTRACT

In this paper, a new text-independent speaker recognition method is proposed. This method uses a modeling of the spectral evolution of the speech signals, which is capable of processing some aspects of the inter-speaker variability : the AR-Vector models. Some inter-speaker measures are presented and their advantages/inconvenients are discussed. A training technique to learn discriminant AR-Vector models is proposed.

The evaluation of this method is carried out on the TIMIT database recorded by cooperative speakers without any impostor. A series of text-independent speaker identification experiments are described. There is no specific corpus for the training sentences and the training corpus is different from the test corpus. Two speech qualities are tested (i.e., good quality and phone quality). The experiments with good speech quality give first-rate results (i.e, identification rate of 100% for 420 speakers) without using more than two sentences for each test.

10.21437/ICSLP.1992-179

I. INTRODUCTION

Speaker recognition refers to three different problems which are the verification of the claimed identity, the speaker identification and the detection of transitions between speakers during the talk. In this paper we are interested in speaker identification which is the most difficult of these three problems. We don't used specific sentences in the training step in order that the results don't depend on the phonetic contents of the identification sentence. The identification task consequently becomes more difficult, but it is more adapted to a real application.

We develope an original speaker-independent recognition system based on the use of a model of the spectral evolution, the AR-Vector Models (ARVM) , and on the use of a new inter-speakers measure we called Itakura-Vector Measure (IVM). The ARVM allows for the description of the spectral parameters trajectories of the speech segment (i.e., sentences or series of sentences).

For the training and the identification step, two modes are described : a discriminant mode and a non-discriminant mode. We introduce, for the dicriminant training, the Discriminant AR-Vector Models (DARVM). The discriminant identification uses a Discriminant Itakura-Vector Measure (DIVM). The discrimant mode gives the best results but it's more complex.

The training step consists of the computation of a model *ARVM (i.e., ARVM or DARVM) for each speaker. The identified speaker is the one of which the training model is the nearest in the sense of the measure *IVM (i.e, IVM or DIVM). For a new speaker, it's only required to compute a new model.

II. AR-VECTOR MODELS

The AR-Vector modeling is a classical tool used to process a signal with various components. It is used here to describe the trajectories of the vectors analysis of speech.

Let $\{y_n\}$, (n = 1 ,.., N) be a succession of N spectral vectors of order p. Their evolution is described by an auto-regressive vectorial model of order q.

$$y_n = \sum_{i=1}^{q} A_i\, y_{n-i} + e_n$$

where $\{A_i\}$, (i = 1 ,.., q) are p x p matrices and $\{e_n\}$, (n = 1 ,.., N) is a vectorial white noise wich has a covariance matrice D. The coefficients of matrices A_i are estimated by the algorithm of Levinson-Whittle-Robinson [1]. The criterion to minimize is the trace of the covariance matrice D. The ARVM models have recently been used with analysis vectors of the speech. Applied to speaker recognition [2][3][4][5][6], they are interpreted as a representation of articulatory capacities of the speaker (i.e., instantaneous mean velocity and instantaneous mean acceleration of spectral parameters). The principal difficulty of this modeling is the estimation of an optimal order q.

2.1. Estimation of optimal order of ARVM model

The optimal order q_{opt} of a ARVM model is difficult to estimate. We choose to use the Akaike's criterion [7], though this criterion has to be applied for monodimensional gaussian signals. We choose to use an order 2. In our experiments of speaker recognition, we note that for a model of order greater than 2, the prediction error doesn't decrease significantly, but the recognition rates decrease considerably. These results are consistent with other experiments on the speech signal modelling [8]. The interpretation of such an order (e.g., 2 or 3) is that the transition mode of the parameters is rather simple and adequatly approximated by a low order ARVM model.

III. INTER-SPEAKERS MEASURES

Speaker recognition consists in associating an unknown test utterance with the original speaker being among a set of training speakers. For this identification, it's necessary to define a measure between speakers. We present two measures IS_O and IS_I and we develop the computation of a third more discriminant one. The last two measures are developed from the study of the prediction error.

3.1. Study of the prediction error

The study of the prediction error is the way in which we define a performant inter-speakers measure. For instance, on the Figure 2., we note that the principle of speaker identification must be based on likeness between the inverse filtering and the vectorial residual but not on the magnitude of the inverse filtering.

3.2. Inter-speakers non-discriminant measures

The chosen inter-speakers measures are based on the Itakura measure [9]. Let us introduce the following definitions and notations :

$\{x_n\}$ (n=1,..,M) : M spectral vectors of speaker X.
$\{A_i\}$ (i=1,..,q) : ARVM model of order q for $\{x_n\}$.
$\{y_n\}$ (n=1,..,N) : N spectral vectors of speaker Y.
$\{B_i\}$ (i =1,..,q) : ARVM model of order q for $\{y_n\}$.
$\{eXA_n\}$ (n=1,..,M) : the residual of the vectors $\{x_n\}$ filtered by the model $\{A_i\}$.

$$eXA_n = x_n - \sum_{i=1}^{q} A_i\, x_{n-i}$$

$\{eYA_n\}$ (n=1,..,N) : the residual of the vectors $\{y_n\}$ filtered by the model $\{A_i\}$.

$$eYA_n = y_n - \sum_{i=1}^{q} A_i\, y_{n-i}$$

$\{eXB_n\}$ (n=1,..,M) : the residual of the vectors $\{x_n\}$ filtered by the model $\{B_i\}$.

$$eXB_n = x_n - \sum_{i=1}^{q} B_i\, x_{n-i}$$

$\{eYB_n\}$ (n=1,..,N) : the residual of the vectors $\{y_n\}$ filtered by the model $\{Bi\}$.

$$eYB_n = y_n - \sum_{i=1}^{q} B_i\, y_{n-i}$$

D_{XA}, D_{YA}, D_{XB}, D_{YB} : the covariance matrices of these residuals (i.e., $\{eXA_n\}$, $\{eYA_n\}$, $\{eXB_n\}$, $\{eYB_n\}$).

We define two distinct inter-speakers measures from these covariance matrices. The first measure IS_O is the original measure developed by Y. Grenier [2]. The second measure IS_I is a symetrized version of the Itakura measure. It is based on the fact that the estimation of a model ARVM on a short sentence is worse than on a long sentence. This measure gives better results than the original measure.

Let G(D) be a function of the eigen values of the matrice D. G(D) is a measure of the likeness between D and the zero matrice. For instance, we can use Trace(D) (i.e. sum of the eigen values) or Det(D) (i.e., product of the eigen values). These two functions give the same results, but the computation of the second is faster.

$IS_O(X,Y) = G(D_{YA})$

$IS_I(X,Y) = M * G(D_{YA} * D_{XA\text{-}1}) + N * G(D_{XB} * D_{YB\text{-}1}) / (M + N)$

We note that the computation cost of one computation measure is proportional to the power 3 of the p dimension of the analysis vectors. But we have noticed than the identification rate increases when the p value increases.

The discrimination between speakers doesn't appear in these measures. Each measure looks for the minimisation of the measure between a speaker and his own model and doesn't look for the maximisation of the measure between the others speakers and this model. Then, we define a deviation DIVM between two speakers during the training. This deviation allows to compute the discrimination obtained on the training set. A new measure inter-speakers DIVM improves this discrimination.

3.3. Discrimination of measure

The IVMM deviation between two speakers S_1 and S_2 represents the average of the IVM measures between each training sentences of speaker S_1 and the model trained with the concatenated sentences of speaker S_2.

Let us define :

Ns : number of the speaker of the training database.

Np : number of the training sentence per speaker.

$\{x_{njs}\}$ (n=1,..,N_{js}) : N_{js} spectral vectors corresponding to the j^{th} training sentences of speaker Y_s.

$\{y_{ns}\}$ (n=1,..,M_s) : M_s spectral concatenated vectors of speaker Y_s, with $M_s = \sum_{j=1}^{Np} N_{js}$.

$\{z_n\}$ (n=1,..,P) : P spectral vectors corresponding to the test sentences of an unknown speaker Z.

$\{A_{ijs}\}$ (i=1,..,q)(j=1,..,Np)(l=1,..,Ns) : ARVM model of order q for $\{x_{njs}\}$

$\{B_{is}\}$ (i=1,..,q)(s=1,..,Ns) : ARVM model of order q for $\{y_{ns}\}$

$\{e1_{njs}\}$ (n=1,..,N_{js}) : vector residual $\{x_{njs}\}$ filtered by the model $\{A_{ijs}\}$.

$\{e2_{njS1S2}\}$ (n=1,..,N_{js}) : vector residual $\{x_{njS1}\}$ filtered by the model $\{B_{iS2}\}$.

$\{e3_{njS1S2}\}$ (n=1,..,M_S) : vector residual $\{y_{nS1}\}$ filtered by the model $\{A_{ijS2}\}$.

$\{e4_{nS1S2}\}$ (n=1,..,M_S) : vector residual $\{y_{nLS}\}$ filtered by the model $\{B_{iS2}\}$.

$D1_{jS}$, $D2_{jS1S2}$, $D3_{jS1S2}$, $D4_{S1S2}$,: covariance matrices of the vector residuals (i.e., $\{e1_{njS}\}$,$\{e2_{njS1S2}\}$,$\{e3_{njS1S2}\}$, $\{e4_{nS1S2}\}$).

IS_{IjS1S2} : The measure IS_I between the j^{th} training sentence of the speaker S_1 and the model of the speaker S_2.

$IS_{IjS1S2} = (N_{jS1}\, G(D2_{jS1S2}\, D1_{jS}^{-1}) + G(D3_{jS1S2}\, DS_{S1S2}^{-1}))\,/\,(N_{jS1} + M_S)$

The IVMM deviation between a speaker S_1 and a speaker S_2 of the training set, is defined by :

$$IVMM(S_1, S_2) = \sum_{j=1}^{Np} IS_{IjS1S2} / Np.$$

The intra-speaker IVMM deviations equals zero when the number of training sentences Np equals 1. It measures the auto-coherence of training sentences. Knowledge of these deviations allows to develop a discriminant measure using the measure IS_I.

3.4 Discriminant inter-speakers measure

For a discriminant measure DIVM, the properties of the IVMM deviation computed with this measure are the following : the IVMM intra-speaker deviation must be small and the inter-speakers deviation must be high. The chosen solution is to compute DIVM as a linear combination of measures $\{IS_I(Y_S,Z)\}$ (S=1,..,Ns) between the test sentence and the speakers models.

$$IVMM(Y_S,Z) = \sum_{k=1}^{Ns} F_{Sk}\, IS_I(Y_k,Z)$$

The matrice F (Ns x Ns) of coefficients F_{Sk} is computed from the matrice M (Ns x Ns) of the deviations ($M_{S1S2} = IVMM(S_1, S_2)$).

$F = J\, M^{-1}$, with $J_{S1S2} = M_{S1S2}$ if $S_1 \neq S_2$,0 otherwise.

The computation cost of the measure DIVM becomes high when the speaker number Ns increases. To decrease this cost we don't compute the matrice F on the whole speaker set, but on every test on the k-nearest speakers of the test sentence. This measure allows a significant improvement of the identification rate. The computation cost remains important. It is better to make a discriminant analysis at the training step (i.e., on the ARVM speaker models) than making a discriminant analysis at the test step (i.e., on the IVMM inter-speakers measures) . To this end, Discriminant AR-Vectors Models (DARVM) have to be computed.

IV. DISCRIMINANT AR-VECTOR MODELS

The DARVM models are defined by the ARVM models minimizing the ratio between the mean of the intra-speaker deviation and the mean of the inter-speakers deviation. The mathematical solution shows that a minimum doesn't always exist. Moreover the compution cost of a numeric solution (e.g., gradient method) is proportional to the power 5 of the ARVM model order p. For these two reasons, we prefere an discriminant analysis of analysis vectors to improve the discrimination of the ARVM models.

4.1 Linear Discriminant Analysis

We look for a projection of the analysis vectors of p order in a sub-space of order r which minimizes the intra-speaker analysis vectors variances and maximizes the inter-speakers analysis vectors variances [10].

Let us define :

Ns : number of the speaker of the training database.

Np : number of the training sentences per speaker.

$\{x_{njs}\}$ (n=1,..,N_{js}) : N_{js} spectral vectors corresponding to the j^{th} training sentences of speaker Y_s.

$\{y_{ns}\}$ (n=1,..,M_s) : M_s concatenated spectral vectors of speaker Y_s, with $M_s = \sum_{j=1}^{Np} N_{js}$.

Av_{js} : average of the N_{js} spectral vectors corresponding to the j^{th} training sentences of speaker Y_s.

μ : average of the Av_{js} for (j=1,..,Np) and (S=1,..,Ns).

μ_s : average of the Av_{js} of the speaker S, for (j=1,..,Np).

K_S : covariance matrice of the analysis vectors of the speaker Y_s.

$$K_S = \frac{1}{Np} \sum_{j=1}^{Np} (Av_{jS} - \mu_S)(Av_{jS} - \mu_S)^t$$

W : intra-speaker analysis vectors covariance matrice.

$$W = \sum_{S=1}^{Ns} \frac{1}{Np} K_S$$

B : inter-speakers analysis vectors covariance matrice.

$$B = \sum_{S=1}^{Ns} \frac{1}{Np} (\mu_S - \mu)(\mu_S - \mu)^t$$

The projection matrices are computed from the r eigen vectors associated to the r highest eigen values of the $W^{-1} * B$ matrice. We hope to decrease the computation cost recognition (i.e, proportional at r^3), without to decrease the identification rate.

V. DATABASE

The text-independent speaker recognition is characterized by the ability to identify a speaker uttering any test corpus. Moreover, during the training, no specific sentence must be used. The chosen corpus must allow the simulation of this recognition.

We used for our experiments the TIMIT database. A full description of this database can be found in [11]. It consists in recording 420 speakers (i.e., 130 females and 290 males). The speakers are categorised into one of the eight "dialect regions" that approximatively map the speech dialects in American English.

Two qualities of speech are tested in these experiments. The first quality corresponds to the original TIMIT database (i.e., 16 kHz and 16 bits). The second quality results of a filtering (0-3400Hz) and under-sampling of the TIMIT database (i.e., 8 kHz and PCM code). This is a simulation of the phone quality of speech.

Each speaker utters 10 sentences. 2 of these sentences are "dialect sentences" that are uttered by every speaker, the 8 other sentences are different for each speaker. 5 of them are "MIT" sentences, the remaining 3 are "TI" sentences. The "MIT" sentences (i.e., 450 sentences) are designed to provide a rich variety of phonetic segments and phonetic contexts. The "TI" sentences (i.e., 1890 sentences) are taken from a large corpus of written text. We note there are no impostor only cooperative speakers and one recording session per speaker. Consequently, the study of the temporal drift isn't possible.

VI. EXPERIMENTS

The 5 "MIT" sentences are used for the training of a model *ARVM (i.e., ARVM or DARVM). For each speaker, the training model is computed from the representation of the concatenation of the corresponding training sentences. The test database consists of the 5 remaining sentences. The speaker identity is determined from one or two sentences.

The chosen parameters are the LPCC (Linear Prediction Cepstral Coefficients). For the experiments 20 coefficients are used.

To identify a speaker uttering a test sentence, the test model is computed. The identified speaker will be the one of which the training model is the nearest in the meaning of the measure *IVM (i.e., IVM or DIVM) of the test model.

6.1. ARVM Models and IVM Measures

For each IVM measures (i.e., IS_O and IS_I) the results of the whole database are given (i.e., identification rate on 420 speakers). Indicative results are given for the N speakers recognition. These results correspond to an average of 100 recognition rates obtained from 100 sub-corpus of N speakers fired at random among the 420 speakers.

Speakers Number	Speech Quality	20	50	100	250	420
IS_O	Good	99.1%	98.1%	97.4%	96.2%	95.3%
	Phone	93.2%	88.6%	84.6%	78.0%	75.0%
IS_I	Good	99.8%	99.5%	99.3%	98.7%	98.4%
	Phone	95.5%	92.2%	89.1%	83.3%	81.3%

Table 1. - Identification on N speakers with one test sentence

Speakers Number	Speech Quality	20	50	100	250	420
IS_O	Good	99.9%	99.7%	99.6%	99.2%	98.9%
	Phone	97.7%	95.9%	94.1%	90.6%	89.6%
IS_I	Good	100%	100%	99.9%	99.8%	99.8%
	Phone	99.0%	98.0%	97.0%	94.9%	94.0%

Table 2. - Identification on N speakers with two concatenated test sentences

We remark (cf. table 1 & 2) that the identification rate on the one hand decreases when the speaker number increases, on the other hand increases when the length of the test increases. The speech phone quality degrades the results, but these are however excellent.

6.2. ARVM models and DIVM measure

The identification rate on the 420 speakers is given in function of the chosen number k (i.e., k-nearest neighbour) introduced for the computation of the matrice F.

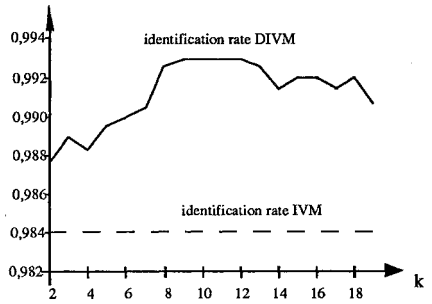

Fig. 1. - Identification rate of DIVM measure in function of k for one test sentence and good speech quality.

We note (cf. Figure 1.) that the identification rate with the DIVM measure, is always superior to the rate obtained by the IVM measure (IS_I). This identification rate has a maximum (99.3%) about k = 10 (i.e., 10-ppv). The following table gives the identification rate with the DIVM measure.

	One sentence		Two sentences	
Speech Quality	Good	Phone	Good	Phone
Identification Rate	99.3%	84.7%	**100 %**	96.3%

Table 3 - Identification rate versus the speech quality and the number of concatenated test sentences for the DIVM measure.

We note a first class result for the good quality speech experiment. There is no error for the identification on two concatenated sentences on 4200 tests.

6.3. DARVM models et DIVM measure

For the DARVM models experiments, a discriminant analysis is used on the analysis vectors. For these experiments the chosen r projection space size is 8. To estimate the contribution of the discriminant analysis, we compare the results of the DARVM model of the analysis vectors projection to the ARVM model of 8 LPCC analysis vectors. The results are significatively better with the DARVM model (90.3%) than the ARVM model (88.5%) for the same number of coefficients. But we can note that these results are lesser than the results of the ARVM model of 20 LPCC analysis vectors (99.3%). The advantage of this method is the identification rate improvement for the same number of coefficients (i.e., for the same cost computation).

VII. CONCLUSIONS

The results show undoubtedly that the ARVM models allow the representation of an important part of the speaker characteristics. Speaker recognition, based on an order 2 ARVM model, implicitely uses the instantaneous mean velocity and the instantaneous mean acceleration of spectral parameters which are specific speaker characteristics (e.g., the elocution speaker rate). We now have to test the speaker recognition system in real conditions with non-cooperative speakers and impostors.

The ARVM models represent the effects of speaker charateristics on the spectral evolution. The obtained results are very encouraging for the search of speaker normalization techniques on the analysis parameters. For instance, such normalization techniques are able to use the residual part of the spectral evolution. It can be used as an independent speaker recognition system or as the voice conversion in a speech synthesis system.

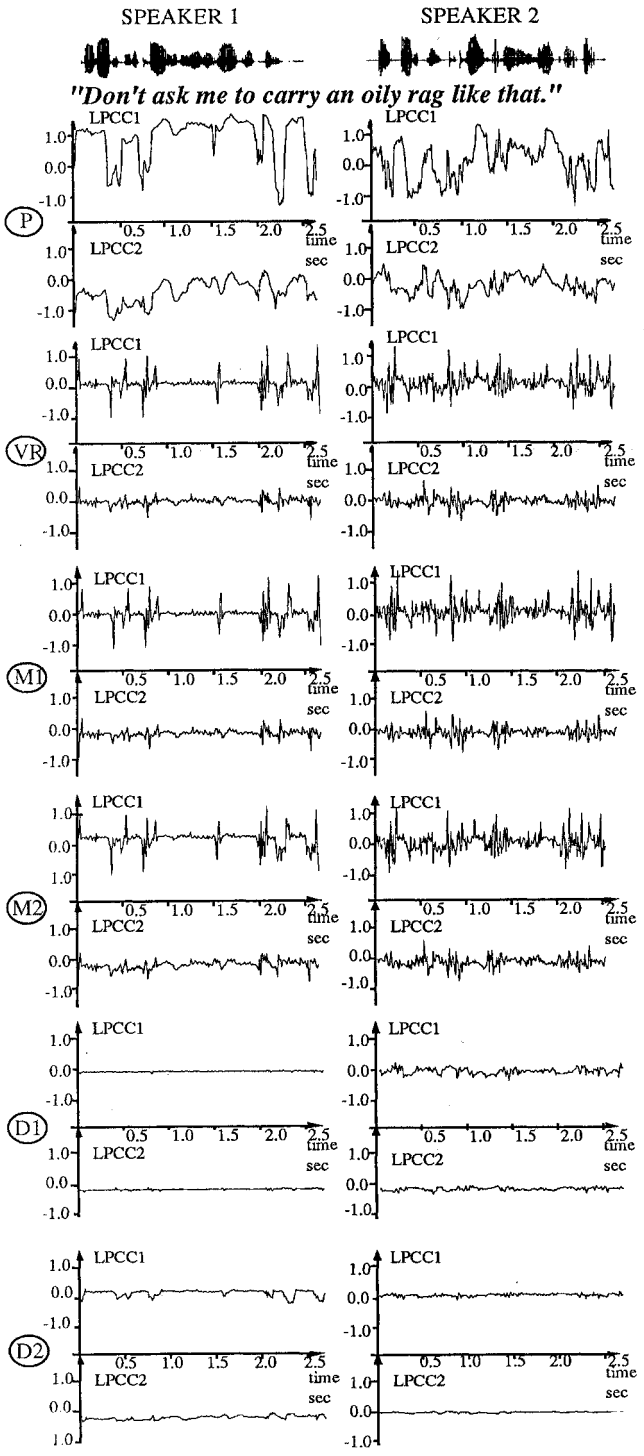

Fig. 2. - Study of the prediction error (the speakers models are trained from different sentences of the test sentence)
P : Evolution of the first 2 LPCC
VR : Vectorial residual of the analysis vector P
M1 : Inverse filtering by the model of the speaker 1
M2 : Inverse filtering by the model of the speaker 2
D1 : Difference between M1 and the residual VR
D2 : Difference between M2 and the residual VR

REFERENCES

[1] P. WHITTLE : "On the Fitting of Multivariate Autoregression and the Approximate Canonical Factorization of a Spectral Density Matrix." Biometrika, Vol. 50, pp. 129-134, 1963.

[2] Y. GRENIER : "Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur." XIème JEP, pp. 163-171, Strasbourg, 1980.

[3] T. ARTIERES, Y. BENNANI, P. GALLINARI & C. MONTACIE : "Connectionist and Conventional Models for Free-Text Talker Identification Tasks." Neuronimes, Nimes, 1991.

[4] C.MONTACIE, P. DELEGLISE, F. BIMBOT & M.-J. CARATY : "Cinematic Techniques for Speech Processing : Temporal Decomposition and Multivariate Linear Prediction" IEEE-ICASSP, San Francisco, 1992.

[5] F. BIMBOT, L. MATHAN, A. DE LIMA & G. CHOLLET : "Standard and Target driven AR-Vector Models for Speech Analysis and Speaker Recognition." IEEE-ICASSP, San Francisco, 1992.

[6] C.MONTACIE, J.L. LE FLOCH & X. RODET : "Modéles autorégressifs vectoriels et reconnaissance du locuteur" 19ème JEP, pp.439-443, Bruxelles, 1992.

[7] H. AKAIKE : Information Theory and an Extension of the Maximum Likehood Principle. 2nd Int. Symp. on Informatic Theory. Tsakhadsor, Arménie, URSS, 1971.

[8] O. KAKUSHO & M. YANAGIDA : "Hierarchical AR model for Time Varying Speech Signals." IEEE-ICASSP., pp. 1295-1298, Paris, 1982.

[9] F. ITAKURA : "Minimum Prediction Residual Principle Applied to Speech Recognition." IEEE Trans. ASSP, Vol. 23, pp. 67-72, 1975.

[10] W.-R. KLECKA : Discrimination Analysis, SPSS,ed by NIE N.H. and al.,pp. 434-467, 1975.

[11] W. FISHER , V. ZUE, J. BERNSTEIN & D. PALLET : "An Acoustic-Phonetic Data Base." J. Acoust. Soc. Amer. Suppl. (A), 81, S92, 1986.