A STUDY ON VOWEL BEHAVIOUR

AND ITS DESCRIPTION BY A STATISTICAL MODEL

M.Domenica Di Benedetto (*) M.Gabriella Di Benedetto (**)

(*) Dpt.of Systems and Computer Sciences (DIS) University of Rome -ITALY
(**)Dpt.of Information and Communication (INFOCOM) University of Rome -ITALY

ABSTRACT

A complete statistical model for italian vowels is presented. The classification method contains the adaptive procedure which makes possible the evolution of the regions characteristic of each vowel in the F1-F2 plane and the effectiveness of which has been verified in a previous work [1]. In this paper, each part of the algorithm is described in its theoretical detail and justified on the basis of an extended experimentation. Results are given proving the accuracy of the classifier for vowel segments independent of the context and prosody and dependent upon the context and prosody.

INTRODUCTION

The problem of eliminating inter-speaker differences without using any a priori knowledge about the input speaker has been analyzed in a previous paper [1]. A method has been proposed as an alternative to the normalization of the recognition parameters by computing the vocal-tract length [2]. This method makes possible the evolution of the characteristic regions of each vowel (which we call "vowel zones") in the first and second formant frequencies (F1 and F2) plane, representing the statistics of several speakers, to the vowel zones of the speaker under examination, representing the statistics of this specific speaker. Further work has been done to improve and complete the model, by investigating the context and prosody dependency of the vowels for a single speaker (for several speakers) and by specifying in the model this component of variability.
The complete statistical model used is illustrated and the updating procedure is justified on the basis of the experimental results obtained.

DESCRIPTION OF THE ALGORITHM

In this section, the algorithm adopted is described in detail by considering:

1. the analysis of the n^{th} frame
2. the classification procedure
3. the adaptive procedure.

As the classification procedure is very simple, more emphasis will be put on the adaptive aspects of the algorithm by carefully investigating its mathematical description and its associated physical meaning.

Analysis of the n^{th} frame

The analysis is made on speech segments of a duration of 12.8 ms. The speech signal is filtered at 5 kHz and sampled at 10 kHz . The quality of the signal considered is good ($\cong$40 dB of S/N ratio). The isolated words as well as the continuous sentences of natural speech used for the experimentation are acquired in the normal experimental environment.

For the analysis of the n^{th} frame, two parameters have been decided to be sufficient information to give to the classifier: the well known first and second formant frequencies. A study has been made to verify the opportunity of using also the third formant frequency; experimental results have shown that the improvement obtained by adding this third parameter is practically inexistent. This fact can be explained by the following consideration: as for the italian language there are only seven classes of vowels (only particular dialects could present more classes of vowels), the corresponding vowel zones have room enough to be sufficiently separated, and their characteristics in the F1-F2 plane are sufficiently different to be represented by non-overlapping ellipses. Problems arise for the vowels /e/ and /ɛ/, as well as for /o/ and /ɔ/, but they are mainly due to the insufficient accuracy with which the manual classification is used in order to generate the initial statistics. In fact, it is often very difficult to determine whether these vowels are open or closed when pronounced in continuous sentences since the speaker does not pay attention to the correct pronounciation of the words, being more influenced by his own way of pronouncing them. This is not the case when isolated words are pronounced (the speaker can be instructed to pronounce the words with the right pronounciation and he "collaborates") and the situation for these vowels is not critical.

More particular attention must be paid to the confusion between /i/ and /e/ and between /u/ and /o/, but also in this case the third formant has not shown to solve these problems.

CH2118-8/85/0000-1557 $1.00 © 1985 IEEE

The classification procedure

The classification procedure is based upon the classical maximum a posteriori probability rule , as already described in [1]. For each vowel segment under examination it is decided that it belongs to class i when one has:

$$P(V_i) > P(V_j) \quad \text{for any } j \neq i$$

where

$$P(V_i) = \frac{P_i\, g_i}{\sum_j P_j\, g_j}$$

P_i indicates the a priori probability and g_i the gaussian probability density function (pdf), corresponding to vowel class j.

The choice of considering gaussian pdf is not restrictive but has been justified in [1] (and consequently the use of elliptic areas as a means for representing the vowel zones).

Updating procedure

The updating procedure has been intuitively described in [1] and here its rigourous justification is presented (see also [3]).

Let $g_i(0)$ be the gaussian pdf , corresponding to the vowel class i, at the first step. This pdf is characterized by the following parameters:

1. the mean value $m_i(0)$
2. the covariance matrix $C_i(0) = F_i + U_i(0)$

As one can see, the covariance matrix is the sum of two matrices F_i and $U_i(0)$,where F_i does not change as the classification proceeds in successive steps, and $U_i(0)$ is updated as soon as a vowel segment is decided to belong to class i.

Updating equations have to be considered now, in order to show how the vowel zones can evolve towards those of the speaker under examination.

At step n, after having analyzed the vowel segment and having attributed it to the vowel class V_i, the parameters of the corresponding gaussian pdf g_i are updated in the following way (by x(n) we indicate the vector of measurements at step n)

$$m_i(n+1) = F_i[F_i+U_i(n)]^{-1} m_i(n) + U_i(n)[F_i+U_i(n)]x(n)$$
$$U_i(n+1) = F_i[F_i+U_i(n)]^{-1} U_i(n)$$

If one writes these equations in relation to step one (n is the number of times that the updating has been applied for class i) one obtains:

$$m_i(n+1) = (F_i[U_i(1)+F_i/n)^{-1} m_i(1) + U_i(1)[U_i(1)+F_i/n]^{-1} . \sum_i x(i))/n$$
$$U_i(n+1) = U_i(1)[U_i(1)+F_i/n]^{-1} . F_i/n$$

When n is large, from these two last equations it results that:

1. $m_i(n)$ tends to $(1/n).\sum_* x(k)$ which is the vector of mean values of the speaker under examination;
2. $U_i(n)$ tends to zero;

When n is large one has then:

3. $C_i(n)$ tends to F_i;

Now, what do the matrices F_i, $U_i(n)$ and $C_i(n)$ represent?

Let us consider two cases separately:

1) recognition of vowel segments which are independent of the context and prosody, called ICP segments (like vowels in a constant and appropriate consonant context in isolated words).

2) recognition of vowel segments dependent upon the context and prosody, called DCP segments (like vowels extracted from natural sentences).

In the first case, the initial gaussian distributions g_i (i=1,7) are characterized by the following parameters:

1. $m_i(0)$: vector of the mean values of F1 and F2 computed on ICP segments and pronounced by several speakers;
2. F_i : covariance matrix which characterizes the mean broadness of the gaussian of a single speaker when ICP segments are considered;
3. $U_i(0)$: covariance matrix of F1 and F2 which takes into account the variations due to the presence of several speakers vowel segments;

When n is large by keeping in mind the previous considerations one has :

1. $m_i(n)$ tends to the mean values of F1 and F2 of the speaker under examination.
2. $U_i(n)$ tends to zero. The variances of F1 and F2 due to the presence of many speakers during the training phase tend to zero .
3. $C_i(n) = F_i + U_i(n)$ tends to F_i. The covariance matrix tends to the one which characterizes the pdf of a general single speaker (F_i represents a "mean" behaviour).

In the second case , the initial gaussian distributions have to contain the information corresponding to the variances of F1 and F2 due to the presence of vowel characteristics variations with the context and prosody. The vector of mean values $m_i(0)$ and the covariance matrix $U_i(0)$ will be the same as before but F_i has to take into account also the variations due to the context and prosody and then is computed on the basis of DCP segments.

EXPERIMENTATION

Training phase

In order to be able to classify ICP and DCP segments a study has been done to obtain the statistics necessary to the classification algorithm. The two cases are described separately.

1 ICP segments :

The analysis has been done for 10 speakers

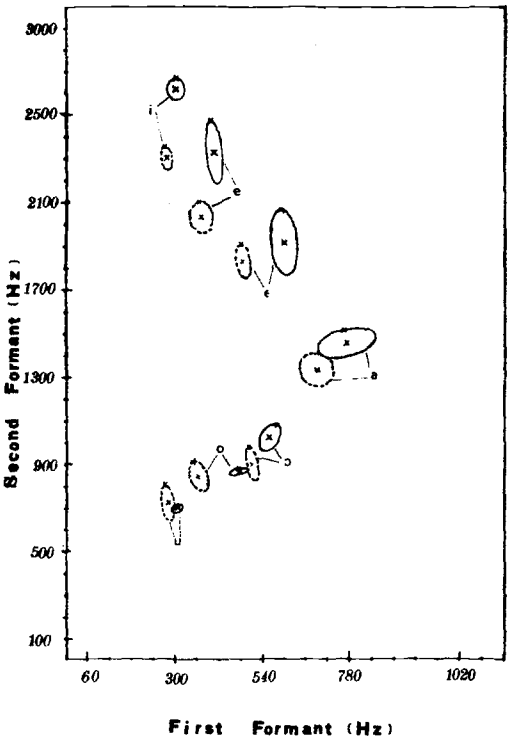

Fig.1. Vowel zones- ICP segments- 1 speaker-

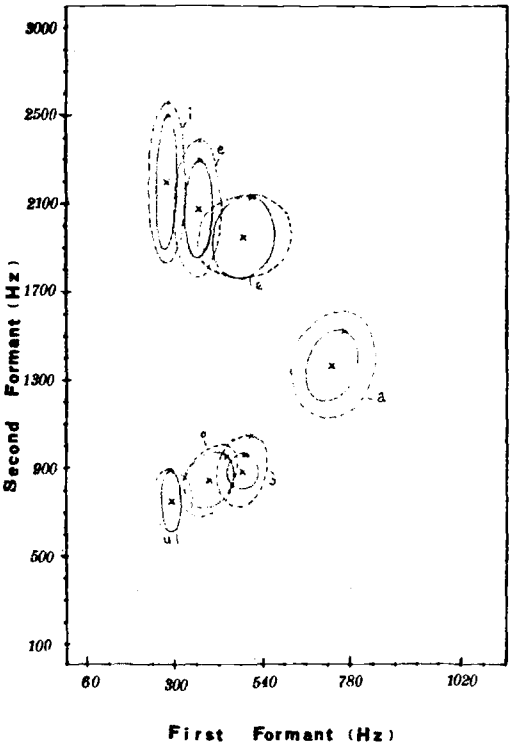

Fig.2. Vowel zones-DCP segments-several speakers

(5 females and 5 males). Vowels segments belonging to isolated words have been considered. Each vowel is pronounced in the same consonant context: /p/a/z/, /p/e/z/, /p/i/z/, etc..

At a first step, for each speaker,the parameters of each pdf have been found and represented in the F1-F2 plane. Fig.1 shows the vowel zones in this case for a male speaker (in dotted line) and the ones for a female speaker (in solid line). The ellipses are narrow as these vowel zones correspond to a single speaker ICP vowel segments.

At a second step, the parameters of the statistics for the population of speakers considered have been estimated. Since at the classification step we want to dispose of initial statistics in which the characteristics of the speaker under examination are not included, the parameters have been estimated on the analysis of segments pronounced by 8 speakers (4 males and 4 females). This operation is repeated 5 times each time with different speakers. By this way, the initial statistics considered never include the characteristics of the speaker under examination (the "test" speaker never belongs to the training set of speakers).

Fig.2 shows an example in this case of the vowel zones (solid line). The ellipses are broad as they take into account the characteristics of several speakers, but not as broad as the ones obtained considering also the variability due to the context and prosody (which will be shown in the next paragraph).
By observing the vowel zones in solid line of Fig.2, one can see that at this step confusions may rise between /o/ and /o/, /e/ and /e/, and that the ellipses become critically near for vowels /i/ and /e/, as well as /u/ and /o/.

2. DCP Segments:

When DCP segments have to be classified, initial statistics are needed which take into consideration all components of variability (speakers, prosody and context).

In order to obtain these data, a study has been done on the "mean" behaviour of parameters F1 and F2 with changes of context and prosody.

For a single speaker, vowel segments belonging to continuous sentences have been analyzed and the results have been compared to those obtained with ICP segments of the same

speaker. Supposing that the variability due to the context and prosody only influences the variances of the distributions (the ellipses are broader but the mean values do not move), it is immediately possible to deduce, by this comparison, the covariance matrices which take into account the variability due only to the context and prosody.

The general statistics, considering all the types of variability are finally obtained from these last data and the statistics of ICP segments. Fig.2 shows an example of the vowel zones in this case (in dotted line). It is not surprising that these ellipses are broader than the ones in solid line as also the context and prosody are considered now.

Test phase

In [1], results have been given on the capability of the updating procedure to enable vowel zones evolution.

ICP SEGMENTS	a	ɛ	e	i	o	ɔ	u
% correct classif.	100	98	95	95	95	98	97

Tab.1. Classification results- ICP segments

DCP SEGMENTS	a	e	ɛ	i	o	ɔ	u
% cor. clas. (stable part)	98	92	94	92	87	90	93
% cor. clas. (trans.incl.)	95	92	85	90	87	90	91

Tab.2. Classification results- DCP segments

In this paper results are presented on the accuracy of the recognition algorithm in classifying vowel segments in a constant and appropriate context and contained in natural sentences.

Table 1 shows the classification rates for vowel segments belonging to the ICP class.

Table 2 illustrates the results obtained in the case of DCP segments extracted from the stable zone of the vowel, and for DCP segments containing also transitions and more ambiguous segments.

It is important to note that the classification becomes more precise (the classification rate improves) when the number of segments analyzed increases, because of the adaptation of the vowel zones to those of the speaker under examination (evolution of the vowel zones towards a non-overlapping situation).

In the case of ICP segments, the number of segments of which we dispose is not high: therefore, the adaptation may not be complete, but it should be noted that the initial statistics are not critical (see Fig.2 in solid line).

In both cases (ICP and DCP segments) the classification rates are satisfactory.

CONCLUSIONS

A method has been presented which allows the classification of vowel segments in a constant consonant context as well as contained in natural speech, without any a priori knowledge about the speaker and any type of normalization.

Experimental results have shown that in both cases the method presented is highly efficient.

The authors understand that the problems that arise for italian vowels are fewer than for other languages since the number of vowels is low and the possibilty of overlapping between the vowel zones is less frequent. Consequently , the conclusions on the effectivness of the algorithm obtained by our experiments cannot be immediately extended to other languages for which this method could prove to be less appropriate than the normalization technique .

The application of this method to other languages could be a field for further interesting investigations.

REFERENCES

[1] M.G.Di Benedetto, A.Lanaro, "How to avoid vowel normalization in the identification of vowels in contiunuous speech", IEEE Int. Conf.Acoust.,Speech, Sig.Proc., 1983.

[2] H Wakita,"Normalization of voweis by vocal-tract length and its application to vowel identification" IEEE Trans. ASSP, Vol.ASSP-25,No.2,April 1977,pp.183-192.

[3] G.Bruno,M.D.Di Benedetto, M.G.Di Benedetto, A.Cilio,P.Mandarini,"New results on a baysian-adaptive V/UV/S classifier for speech signals" Istituto di Comunicazioni Elettriche, Universita' degli Studi di Roma 'La Sapienza', Internal Report, April 1982.