HOW TO AVOID VOWEL NORMALIZATION IN IDENTIFICATION OF VOWELS IN CONTINUOUS SPEECH

M.Gabriella Di Benedetto (*) Armando Lanaro (**)

(*)Dpt.of Information and Communication (INFOCOM) University of Rome -ITALY

(**)Dpt.of Physics and Astronomy University of Rochester -USA

ABSTRACT

The examination of the diagrams showing the unnormalized formant frequencies in the F_1-F_2 (first and second formant) plane of the vowels (in italian or other languages) shows an overlap of the regions characteristic of each vowel (vowel zones) caused mainly by the differences in the measurement and shape of the vocal tract among speakers. In this paper, a method is presented that makes possible the evolution of the vowel zones by using an adaptive procedure. At the beginning of the algorithm, the center of gravity of each vowel zone represents the mean characteristics for many speakers, when pronouncing each vowel. In the successive steps, the centers of gravity and the shape of the zones tend towards the ones characteristic of the specific speaker. Results are given that show this evolution both for male and female speakers, with the elimination of the overlap between the zones.

INTRODUCTION

The main difficulty encountered in automatic vowel identification in continuous speech is to eliminate inter-speaker differences without using any a priori knowledge about the input speaker. Various techniques were proposed in the past to overcome this problem. One possible method consists in normalizing the parameters used at recognition by computing the vocal-tract length [1]. Another technique, which avoids normalization, is based upon a pattern recognition approach: the vowel under examination is compared with reference patterns spoken by different speakers and stored during a learning phase [2].

When using the unnormalized first and second formant frequencies (F_1 and F_2) as parameters, inter-speaker differences can be pinpointed by examining the overlap of the areas which are characteristic of each vowel (further called "vowel zones") in the F_1-F_2 plane. The method proposed makes the evolution of these zones possible by introducing the concept of adaptability in the vowel recognition system. In this way, the vowel zones tend towards a non overlapping situation.

Work carried out at Fondazione Ugo Bordoni (Viale Europa 160, Rome, Italy).

VOWEL ZONES REPRESENTATION

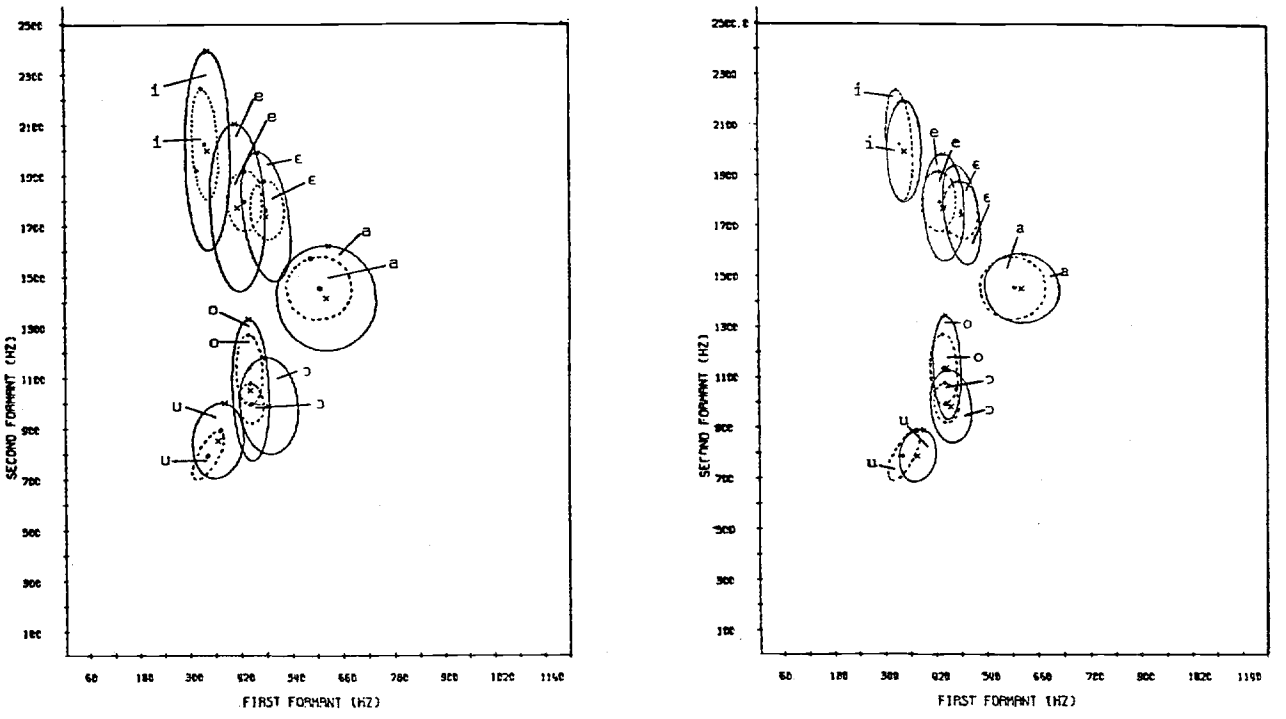
Consider the intersection between a gaussian probability density function (pdf) of the parameters F_1 and F_2 and a plane (Z) parallel to the F_1-F_2 plane. The projection of this intersection on the F_1-F_2 plane is an ellipse the center of which is the mean of the first two formant frequencies; the lengths of the major and minor axes are proportional to the variances. The orientation of the axes depends upon the correlation between the two parameters. Bearing in mind the definition of the pdf, and having chosen a certain probability P, these ellipses can be thought as loci of equiprobability P, when the height of the plane (Z) has been appropriately determined. Having fixed a probability P (and consequently the height of the plane (Z)) the probability of being "inside the ellipse" is P, while the probability of being "outside the ellipse" is 1-P.

The method proposed in this paper will make use of these elliptic areas of the F_1-F_2 plane as representations of the vowel zones. This choice can be justified by the study carried out by Ferrero [3] for the italian language, in which the vowel zones present an elliptic shape. This behaviour is also present for vowels of other languages ([1], [3]).
This type of representation will be used for the vowel zones of several speakers as well as of a single speaker, as both cases will be considered from a statistical point of view (see for example Fig.1 (a)). In this way, for a single speaker, the possible variations in the measurements of F_1 and F_2 due to the variations of the latter with context and to eventual errors and approximations of the formant extraction algorithm are taken into account. For several speakers, this representation will be useful to show the amount of variability of the F_1 and F_2 parameters among speakers.

As the overlapping is due to the inter-speaker differences, the ellipses of a single speaker should result separated and those of several speakers should show some areas of superposition.

THE CLASSIFIER

Consider a gaussian distribution for the



(a) (b)

Fig.1. Vowel zones before (a) and after (b) the adaptation- Male speaker-

F_1 and F_2 parameters for several speakers. The center of gravity of each vowel zone represents the mean characteristics for many speakers when pronouncing each vowel.

The algorithm bases its decision on the maximum a posteriori probability by classifying a vowel segment under examination as belonging to vowel class j when one has:

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

where

$$P(V_j) = \frac{P_j\, g_j}{\sum_i P_i\, g_i}$$

P_i indicates the a priori probability and g_i the gaussian pdf, corresponding to vowel class i.

In order to begin with the classification, it is necessary to give the classifier the initial data it needs for each vowel class, which are the parameters of the statistics for many speakers. The latter are obtained during a learning phase described in section "Statistics".

In the successive steps of the algorithm, the mean and the covariance matrix of the pdf of the parameters are updated in correspondence to the vowel class to which the segment under examination has been decided to belong. To reduce the eventuality of updating statistics corresponding to misclassifications, this operation is done only if the decided vowel class probability is greater than a threshold.

The updating procedure can be intuitively described as follows.

At step n the mean of the pdf to be considered will be obtained as a linear combination of the mean at the previous step and the measurements at step n. The coefficients of this linear combination are such that they attribute to the terms of the combination weights which are inversely proportional to their variances. In this way, a term "weights more" when its variances have become smaller (the distribution has shrunk around the mean value).
The covariance matrix of the pdf will be updated in such a way that it will tend to a constant covariance matrix that takes into account the variability of the measurements of the single speaker.

In this way, the algorithm converges to the specific situation of the speaker under examination. The rigourous updating formulas with their theoretical justifications can be found in [4].

Consider the representation on the F_1-F_2 plane. At the initial state of the algorithm, the center of gravity of each ellipse (solid line in Fig.1 (a) or Fig.2 (a)) represents the mean characteristics for many speakers when pronouncing the corresponding vowel. In the following steps, as the algorithm makes the classification (the classifier acquires information about the speaker), the position of each center of gravity moves towards a point which represents the mean of the specific speaker, while also the shape of each ellipse tends to conform with that of the specific speaker (see Fig.1 (b) and Fig.2 (b)).

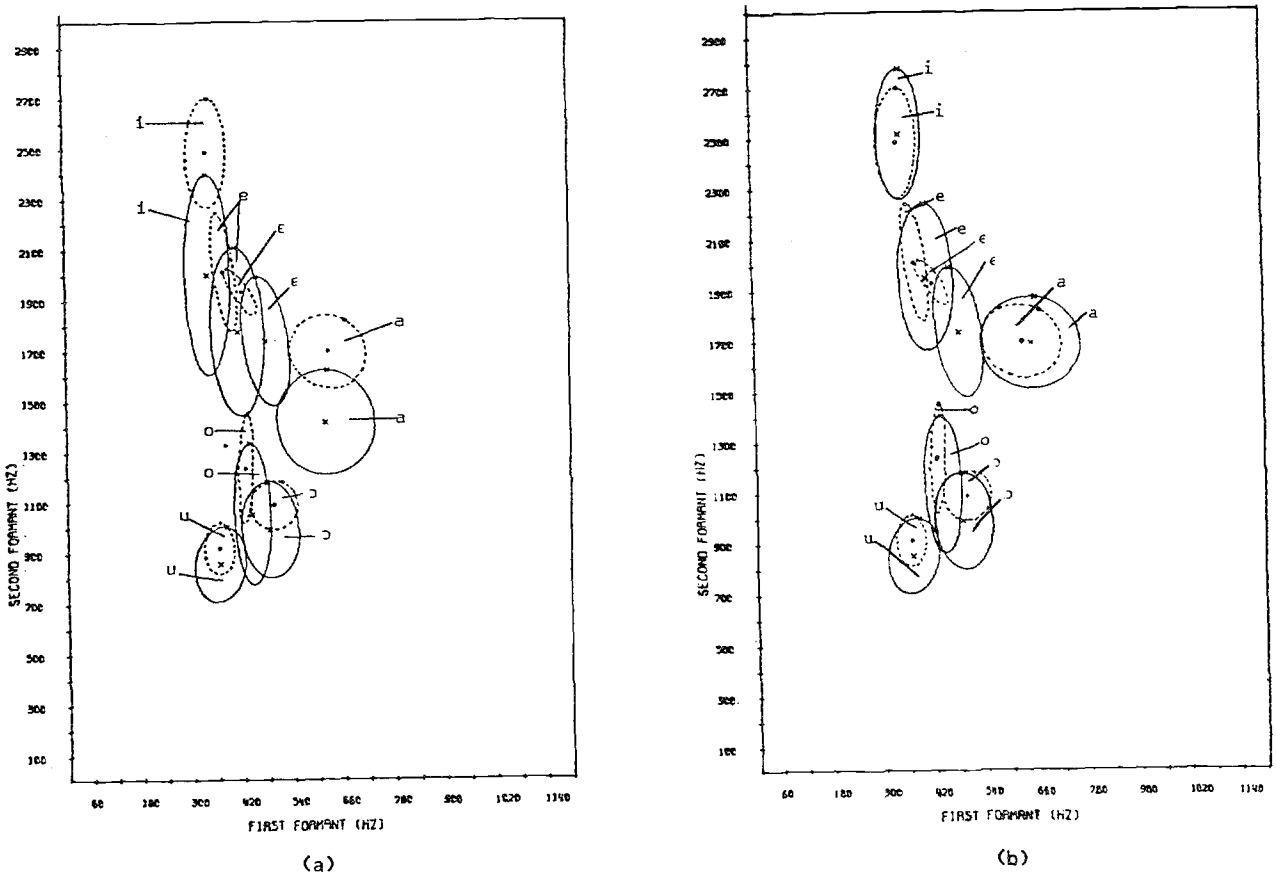


Fig.2. Vowel zones before (a) and after (b) the adaptation-Female speaker-

STATISTICS

The parameters of the statistics necessary at the initial step of the algorithm have been estimated on the basis of vowel segments belonging to the seven italian vowel classes which have been extracted from continuous sentences manually by the examination of the corresponding spectrograms. During this learning phase, two male speakers were considered, each pronouncing three natural sentences. Only two male speakers have been included in the training phase in order to have the most difficult situation when applying the algorithm for example to a female voice, for which the values of the formant frequencies are markly differentiated from those of male speakers.

The vowel zones obtained for the initial pdf that has been considered are shown in Fig.1 (a) or Fig.2 (a) (in solid line). All the vowel zones presented correspond to a probability P equal to 0.98.

EXPERIMENTATION

The purpose was to automatically identify, by means of the algorithm described in the previous paragraphs, vowel segments belonging to continuous sentences spoken by one male and one female speakers not considered in the training phase.

In order to make this possible, an algorithm (briefly described in the next paragraph) has been studied to automatically identify the presence of a vowel segment in running speech. A more detailed discussion of this algorithm can be found in [5].

Vowel Presence Identification

The vowels examined and classified by means of the algorithm described were extracted from continuous sentences by means of a deterministic method based upon the analysis of the frequency distribution of the energy of the speech signal. 14 parameters were considered all given by a percentage of energy in a given frequency band with respect to the energy of another frequency band, or given by the amount of energy in an opportunely chosen frequency band. The decision algorithm operates as a sequence of comparisons with established thresholds. With the quality of speech signal considered (sampling rate of 10 kHz and S/N less than 50 dB) this algorithm has

shown to behave in a satisfactory way, fast and accurate enough in detecting the presence of vowel portions (even of some non steady-state vowels).

Results

The algorithm proposed for the recognition of vowels has been experimented on four sentences of natural speech spoken in italian, of an approximate duration of 25 seconds, pronounced by two native speakers (one female and one male) not belonging to the training phase.

Fig.1 (a) and Fig.2 (a) show the situation of the vowel zones at the first step of the algorithm (solid line). For the sake of clarity, the real statistics of the speakers under test have been computed and shown in dotted line. It is obvious that if the adaptive algorithm is capable of making the vowel zones evolve to a non overlapping configuration specific to the speaker and corresponding to a very high probability, any classification procedure, based upon the F_1 and F_2 parameters, will be then highly efficient. Therefore, we do not present here results on the accuracy of the recognition algorithm but we show the configuration of the vowel zones after the application of the adaptive procedure described (see Fig.1 (b) and Fig.2 (b), solid line).

During the experimentation, the vowel classes o and ɔ, as well as e and ɛ have been confused, due to the often insufficient accuracy in the manual labeling of the sentences.

Fig.1 (b) shows the state of the vowel zones, for the male speaker, after 50 iterations, which has shown to be the beginning of a steady situation. The vowel zones of the general statistics are well adapted to the statistics of the test speaker as regards variances and mean values. The most evident adaptation has in this case concerned the variances, as the mean values were yet close to the final ones desired, as can be seen in Fig.1 (a). This result is not surprising as the training phase has been carried out only with male voices.

Fig.2 (b) shows the situation of the vowel zones for the female speaker after 35 iterations. It is possible, in this case, to draw more interesting conclusions on the capability of adaptation of the algorithm with regards to the mean values. In fact, the latter are here at the initial step, markly far from the final desired ones, due to the absence of female voices in the training phase (see Fig.2 (a)). In Fig.2 (b) the situation has greatly and well evolved. As the more interesting vowels were /a/, /i/ and /e/, the other two vowels /o/ and /u/ rarely occurred in the sentences pronounced. This explains why the evolution of the corresponding vowel zones is almost absent.

CONCLUSIONS

A method has been presented that avoids vowel normalization by making possible the evolution of the characteristics of the vowels (represented by the vowel zones) in the F_1-F_2 plane. Results are given that show the effectiveness of the procedure either for male or female speakers. Improvements are certainly possible with further experimentation by considering a larger training set.

Further work has to be done in the study of the context dependency of the vowels of a single speaker in order to specify in the model this component of variability.

ACKNOWLEDGEMENTS

The authors wish to thank Dr.Maria Domenica Di Benedetto and Prof. Paolo Mandarini for the very useful discussions during the progress of this work.

REFERENCES

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

[2] V.N.Gupta, J.K.Bryan, J.N.Gowdy, "Speaker-independent vowel identification in continuous speech" IEEE Int.Conf.Acoust., Speech, Sig.Proc., 1978, pp.546-548.

[3] F.E.Ferrero, "Diagrammi di esistenza delle vocali italiane" Alta Frequenza,Vol.XXXVII, No.1, January 1968, pp.54-58.

[4] G.Bruno,M.D.Di Benedetto, M.G.Di Benedetto, A.Gilio,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.

[5] M.G.Di Benedetto, A.Lanaro,"Individuazione dei fonemi vocalici nel parlato", Atti del Convegno AIA 1983, Torino sept.1983.