VOWEL DISTINCTION ALONG AUDITORY DIMENSIONS: A COMPARISON BETWEEN A STATISTICAL AND A NEURAL CLASSIFIER.

Maria-Gabriella Di Benedetto and Giovanni Flammia

University of Rome 'La Sapienza'- INFOCOM Dept.- Via Eudossiana,18 - 00184 Rome, Italy.

The aim of this study was to verify whether, in Italian, the use of auditory parameters, such as the Bark-transformed formant differences, is more appropriate to represent vowels than the traditional formant values expressed in Hertz. In addition, the performance of a statistical and of a neural net recognizer, based on the input parameters proposed above was compared.
In the first part of the paper spectral measurements and statistical analyses of all the vowels of the Italian vowel system, uttered by 25 male and 11 female speakers are described. In agreement with a model of American English vowels, it is shown that the difference between the first formant and the fundamental frequency (F1-F0) and the difference between successive formants on the Bark scale (F2-F1 and F3-F2) are effective in normalizing male and female spectral differences and in better clustering vowel areas. The results are discussed on the basis of a model of speech articulation, and an experimental theory of speech perception.
The second part of the paper is a comparison of two vowel recognition methods. Based on the results described in part one, each new vowel input to the classifier is represented by only three values (F1-F0, F2-F1, F3-F2, expressed in Bark). The first classifier is based on a statistical approach and is an application of discriminant analysis. The second classifier is based on a neural network approach and is an implementation of a multi-layer perceptron. The classification error rates are compared and discussed in relation to the different underlying assumptions that the two algorithms make on the input data.

I. Introduction

Automatic speech recognition is an application which raises many questions about our knowledge of the phonation and perception of speech. In other applications as well, such as text-to-speech speech synthesis systems, the problem of a lack of understanding in the fundamental mechanisms of speech production and perception leads to difficulties in conceiving systems which can produce pleasant and natural-sounding voice.

The study described in the present paper tries to put some light on an eternal problem for those who have, in a way or the other, to characterize speech segments in terms of acoustic parameters; if the speech segments are vowel segments, the problem is to find acoustic properties of vowels which prove to be both speaker independent (normalization problem) and context independent (coarticulation problem). It is clear that these two problems, are directly related to a similar issue which consists in finding properties in the acoustic vowel waveform which are invariant with respect to speaker, language, and phonetic contexts variations. The present study will focuse on the first of these problems, following an approach by which the information contained in vowels is coded in a way similar to that used by the auditory system.

As well known, when a vowel is produced, the vocal tract can be considered as a sequence of acoustic tubes which resonate at particular frequencies, called formants. The position of these formants depends essentially upon the length and the cross-sectional area of each tube. If the vocal tract is modeled by a linear system, it can be seen that the formant frequencies F1, F2, F3, F4.. are the imaginary parts of the poles of the transfer function of this system. The losses due to the non-rigid walls characteristics and to the non-ideal junctions and interfaces at the extremities of the tract constitute the real parts of the poles. When a vowel is produced, the source signal is formed by the pressure of the air expirated by the lungs and modulated by the opening and closing of the vocal folds. The spectrum of the source, which is a sequence of pulses of triangular shape with period T0, is then characterized by the harmonics of the fundamental frequency F0=1/T0.

The fundamental frequency, F0, is proportional to the tension of the vocal folds. The variation of the tension of the folds is related to the tension of the surrounding muscles and to the movement of the hyoid bone (Honda, 1983), so that the source and the filter are coupled

during the articulation. The acoustic signal is the product of a convolution between the source signal and the the impulse response of the filter (vocal tract). This signal is periodic and in its spectrum the F0 harmonics which are next to the formants are emphasized. Depending upon which vowel is pronounced, the tongue position varies and, consequently, the size of each of the acoustic tubes, the rigidity of the walls, and the tension of the vocal folds are modified, influencing the F0, F1, F2, F3...values. The acoustic model predicts the relative invariance of the formants of the extreme vowels [i,a,u], when, changing the speaker, the dimensions of the vocal tract are varied (Stevens,1972).

A variation of the sound pressure generates the variation of the hydro-mechanic pressure inside the ear. Many auditory nerve fibers on the basilar membrane of the cochlea are excited and output a signal formed by a sequence of electric pulses which is transmitted to the brain. Each fiber has an action similar to that of a non linear low-pass filter centered around a characteristic frequency CF. These filters cover the spectrum according to a scale which is approximately logarithmic, and which has been formalized in the Bark scale. On the basis of physiological experiments (Delgutte,1984), it was shown that in response to vocalic stimuli the electric pulses generated by all the fibers code the position of the formants. In addition, on the basis of perceptual experiments, it was shown that two peaks in the spectrum of a vowel closer than 3-3.4 Bark are integrated in one peak in intermediate position ('spectral center of gravity effect', Chistovich et al., 1979). As regards the first formant F1, perceptual experiments showed that this formant is perceived relatively to F0 (Traunmüller,1981). Later studies (Di Benedetto, 1987) showed that the relation between F1 and F0 may not be a simple linear relation, implying a non-uniform vowel normalization in agreement with Fant (1975).

In the present paper, two methods for classifying vowels are compared. The first method is based on a statistical approach, the second method is based on a neural net approach. The input to both recognizers is a vector X in which the elements represent the spectrum of the acoustic signal in the way it should be coded by the human auditory system, according to a perceptual model presented by Syrdal and Gopal (1986). In section I, the spectral measurements and the statistical analyses carried out on all vowels of the Italian vowel system pronounced 25 male and 11 female speakers are described. Each vowel is represented by the values (F3-F2, F2-F1, F1-F0), expressed in Bark. In section II, the results obtained by the application of a statistical recognizer are analyzed. The statistical classifier uses a linear discriminant analysis. In section III, the results obtained by the use of a neural net based classifier are analyzed. The neural net classifier is an implementation of a 2-layers perceptron. Finally, the results obtained with these two methods are compared and discussed in the conclusion.

Further detail about the experiments described can be found in Flammia (1988).

II. Experimental conditions and procedure

II.1 Speech material

The Italian vowel system consists of seven vowels [i,e,ɛ,a,o,ɔ,u]. Three vowels [i,e,ɛ] are front vowels, while four vowels [a,o,ɔ,u] are non-front vowels. In the present study, the Italian vowels pronounced by 25 male and 11 female speakers were analyzed. These vowels, extracted from the data-base created by Ferrero (1968), were pronounced in pV# syllables by male speakers and in isolation by female speakers. This data-base constitutes a reference point for many acoustic studies of Italian vowels (see for example Disner, 1983). These data were also used in applications such as a text-to-speech synthesis system . For each vowel, a short temporal window was considered (4 periods long, located around the maximum of the signal enveloppe). For each vowel, the fundamental frequency F0 was computed using an algorithm based on the cepstrum of the signal, and the first four formants were found by manual comparison of the local maxima in the Fourier transform (FFT) of the signal and the maxima of the autoregressive analysis (AR) of the spectrum (linear prediction with 16 coefficients found with the autocorrelation algorithm). The comparison between the two spectra was necessary, as, frequently, the peaks in the AR spectrum under 2000 Hz are slightly lower than those found from the examination of the FFT, when F0 is high (this happens for most of the female speakers). The average values of F0, F1, F2, F3 and F4 expressed in Hertz, for each vowel, for the 36 speakers, are shown in Fig.1.

II.2 Statistical analysis

The values of F0, F1, F2, F3, and F4 have been the object of a statistical investigation. In general, it was found that the measurements expressed in Bark tend to reduce the differences between male and female speakers. Table I indicates the Mahalanobis distance between male and female groups, for different vowels, in the case of the F1 vs F2 (in Hz) and (F1-F0)vs (F2-F1) (in Bark) representations (the Mahalanobis distance is the euclidean distance between the average values of the two groups, divided by the variance of each group). One can note, on Table I, that the Mahalanobis distance between the means for male and female speakers decreases if the measurements are expressed in Bark. Qualitatively, distances lower than 10 indicate a significant overlap between groups. Note that except for the vowel [u], the use of the Bark scale normalizes the differences between male and female speakers.

Figure 2 shows the relation between the standard deviation of F1 and the average values of F1-F0, for different vowels. One can notice, from Fig.2, that a correlation exists between these two parameters, and that in particular when F1 is close to F0, F1 is relatively unchanged for different speakers, in agreement with Di Benedetto's findings (1987). In addition, the distance F1-F0 seems to be correlated to the phonological classification of vowels according to vowel height, in agreement with the expriments reported by Traunmüller (1981).

The relations between the standard deviations of F2-F1 and of F3-F2 with respect to the average values of F2-F1 and F3-F2, all expressed in Bark, are shown in Figs.3 and 4, respectively. These figures show that when two formants are far the variability of the distance between the two formants is higher than when the formants are close. In addition, see from Fig.4 that all front vowels verify F3-F2<3 Bark, while all non front vowels have F3-F2>3 Bark.

III. Statistical classifier

Each vowel was represented by the values of F1-F0, F2-F1, and F3-F2, all expressed in Bark. Given a significant sample formed by a number of utterances of the vowels, each new utterance was classified as a particular vowel according to the following procedure .

First, the new utterance is labeled as front or back, according to the value of F3-F2. Secondly, the two vowels which have the lower euclidean distance in terms of F1-F0 and F2-F1 values from the utterance analyzed, were found.

The choice between the two selected vowels was based on a linear discriminant analysis. The linear discriminant analysis determined the linear combination of the inputs which maximized the difference between the average values of the two groups with respect to the variance of each group. The results obtained by the application of this classifier are shown in Table II. These results are satisfactory for the cardinal vowels [i,a,u] and for front vowels, while they are less satisfactory for back vowels.

IV. Neural classifier

Two 2-layers perceptrons, one for front vowels and the other for non fornt vowels, were trained. A similar architecture for Swedish vowels classification was used very recently by Hult (1989). Each node in a perceptron computes the weighted summation of the inputs. The result goes through a sigmoidal threshold device. In this way, each node creates two regions in the input space separated by a hyperplane. The output of each node of a layer is connected to the input to the nodes of the following layer, in order to inhibit or excite their response. In the case of vowels classification, the input to the first layer of N "hidden" nodes was formed by the vector [X1,X2], where X1=F1-F0 and X2=F2-F1. The N outputs of the hidden nodes were connected to the second layer formed by M output nodes, one for each vowel.

The training of the nets was obtained by the use of the back-propagation error technique, which consists in presenting to the net many samples of the vowel to be recognized. In each cycle, the input [X1,X2] and the desired output were presented to the net. The desired output was high (equal to 1) while for all the other nodes it was low (equal to 0). The output of the perceptron was then computed and the weights were modified proportionally to the error using a constant gain g. After each cycle, the training was speeded up by adding a quantity

proportional to the correction made during the preceding cycle, according to a constant value a. During the first cycles, all the outputs were around 0.5. Slowly, after a number of cycles, the outputs tended to the values 0.1 or 0.9.

Different perceptrons were tested each one being characterized by a different number of hidden nodes N, and different initial weights. The perceptron for the front vowels [i,e,ɛ] was correctly trained after 1080 cycles using N=12, g=0.1, and a=0.9, and all the initial weights being small and of random value. Perceptrons with a lower number of hidden nodes tended to distinguish only extreme vowels. The classification rates and confusion matrix obtained after the training are shown in Table III.

The behaviour of the perceptron for non front vowels [a,o,ɔ,u] was less satisfactory due to the significant overlap between the vowels [o,u] and [o,ɔ]. Independently of the number of nodes, the perceptrons tended to confuse more than 25% of the utterances of [o] and [ɔ], as they were in a way cheated during the back-propagation of the error by ambiguous pronounciations of [o,ɔ], in the region of intersection between [u,o] and [o,ɔ]. In a second series of experiments, the two vowel classes [o] and [ɔ] were confused in one class. In this case, the training of the perceptron was correct with N=20, g=0.1, a=0.4, after 1512 cycles. The classification errors and confusion matrix are shown in Table IV.

V. Conclusion

The statistical analyses and the behaviour of the two classification methods verify that the acoustic and perceptual parameters selected are significant for the distinction of Italian vowels. It is important to point out, however, that it is necessary to carry out studies in which the effects of prosody and coarticulation are analyzed.

As regards the acoustic correlates of distinctive phonetic features for Italian vowels, the overlapping between [o,ɔ] shows that the tense/lax dimension is not well represented by the selected parameters. In fact, this category seems to be related to temporal properties of the first formant trajectory (Di Benedetto, 1988). In addition, the overlapping between [u,o] could be related to the fact that F1-F0 does not represent properly vowel height when F1 and F0 are close and F1 is low, confirming what reported in Di Benedetto (1987).

As regards the applications, a normalization between vowels pronounced by male and female speakers was obtained. The comparison between the statistical and the neural classifiers verifies the equivalence of the two methods when the number of input variables is low and the groups can be separated by a linear combination of the input data. The training of the neural net with the back-propagation error algorithm was time consuming and depended upon the number of hidden nodes of the first layer. When the overlapping between two groups is significant (Mahalanobis distance lower than 10), the behaviour of the perceptron is less satisfactory than the behaviour of the linear discriminant analysis. Nevertheless, recent results reported on neural nets indicate that when the number of variables is high (>10) the statistical procedure loose control on data (computational difficulties in the inversion of the covariance matrix) and consequently it is difficult to discriminate among groups on the basis of a linear combination of the inputs.

References

L.A.Chistovich, R.Sheikin and V.V.Lublinskaja (1979), "'Centres of gravity and spectral peaks as the determinants of vowel quality", in *Frontiers of speech communication research,* edited by B.Lindblom and S.Öhman, Academic Press, London, pp.143-157.

B. Delgutte (1984), "Codage de la parole dans le nerf auditif", Thèse de Doctorat d'Etat, Université Pierre et Marie Curie Paris 6, chap.2, pp.33-78

M.G. Di Benedetto (1987), "On vowel height: acoustic and perceptual representation by the fundamental and the first formant frequency", Proceedings of the 11th Int. Cong. of Phonetic Sciences, Tallin (Estonia, USSR), pp.198-201.

M.G. Di Benedetto (1988), "The influence of the first formant frequency trajectory shape on the perception of the tense/lax distinction of American English and Italian vowels", Studi italiani di linguistica teorica ed applicata, anno XVII, no 2-3, pp.289-297

S.F.Disner (1983), "Vowel quality: the relation between universal and language specific factors", Ph.D. Dissertation, University of California, Los Angeles.

C.G.M.Fant (1975), "Non-uniform vowel normalization", STL-QPSR 2-3.

F.Ferrero (1968), "Diagrammi di esistenza delle vocali italiane", Alta Frequenza, Vol 37, No 1, pp.54-58.

G. Flammia (1988), "Classificazione statistica e neurale su base percettiva nel riconoscimento delle vocali italiane", Tesi di Laurea, Università degli Studi di Roma 'La Sapienza'.

K. Honda (1983), "Relationship between pitch control and vowel articulation", in Vocal Fold Physiology: contemporary research and clinical issues, edited by D.B.Bless and J.H.Abbs, College-Hill Press, pp.286-297.

G.Hult (1989), "Some vowel recognition experiments using multilayer perceptrons", STL-QPSR 1, pp.125-130.

K.N.Stevens (1972), "The quantal nature of speech: evidence from articulatory acoustic data", in *Human Communication: a unified view*, edited by P.B.Denes and E.E.David Jr., McGraw-Hill, New York, pp.51-66.

A.K.Syrdal and H.S.Gopal (1986), "A perceptual model of vowel recognition based on the auditory representation of American English vowels", J. Acoust. Soc. Am., Vol 79 No 4, pp.1086-1100.

H.Traunmüller (1981), "Perceptual deimension of vowel openness in vowels", J. Aoust. Soc. Am., Vol 69 No 5, pp.1465-1475.

TABLE I. Mahalanobis distance between male and female groups in the F1 vs. F2 space (F1 and F0 are expressed in Herz), and in the F1-F0 vs. F2-F1 space (F0, F1, and F2 are expressed in Bark)

Vowel	distance (kHz)	distance (Bark)
[i]	14	4.5
[e]	19	4.2
[ɛ]	11.9	2.1
[a]	10.3	2.8
[ɔ]	2.25	0.11
[o]	1.32	0.79
[u]	0.34	1.62

TABLE II. Classification results obtained with the statistical classifier. Confusion matrix and percentage of correct classification .

	[i]	[e]	[ɛ]	[a]	[ɔ]	[o]	[u]	%
[i]	34	2	0	0	0	0	0	94.4
[e]	0	35	1	0	0	0	0	97.2
[ɛ]	0	3	33	0	0	0	0	91.7
[a]	0	0	0	36	0	0	0	100
[ɔ]	0	0	0	1	31	4	0	86.8
[o]	0	0	0	0	4	26	6	72.2
[u]	0	0	0	0	0	4	32	88.9
average:								90.2

TABLE III. Classification results obtained with the neural classifier, for front vowels. Confusion matrix and percentage of correct classification .

	[i]	[e]	[ɛ]	%
[i]	36	0	0	100
[e]	1	34	1	94.4
[ɛ]	0	1	35	97.2
average:				97.2

TABLE IV. Classification results obtained with the neural classifier, for non front vowels. Confusion matrix and percentage of correct classification .

	[u]	[o,ɔ]	[a]	%
[u]	31	5	0	86.1
[o,ɔ]	3	68	1	94.4
[a]	0	0	36	100
weighted average:				93.7

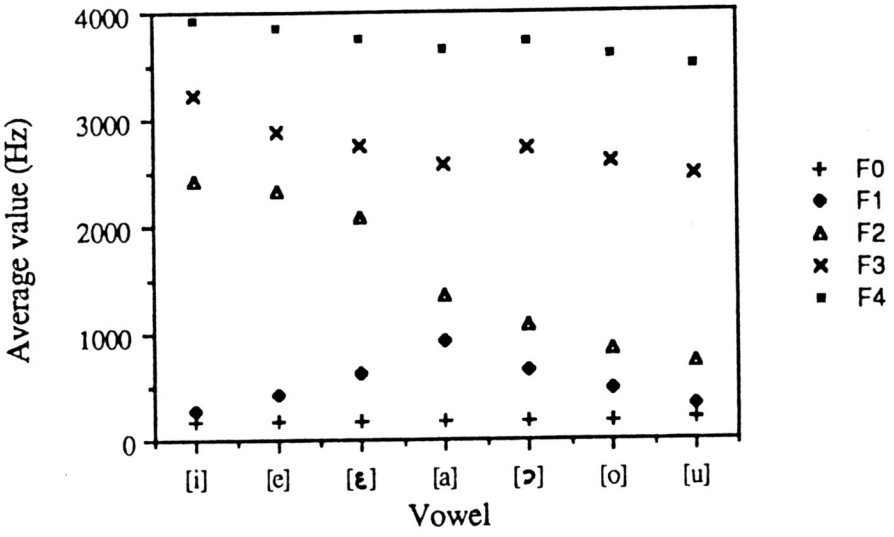

Figure 1 Average values of F0, F1, F2, F3, and F4, for all vowels, considering the 25 male and 11 female speakers.

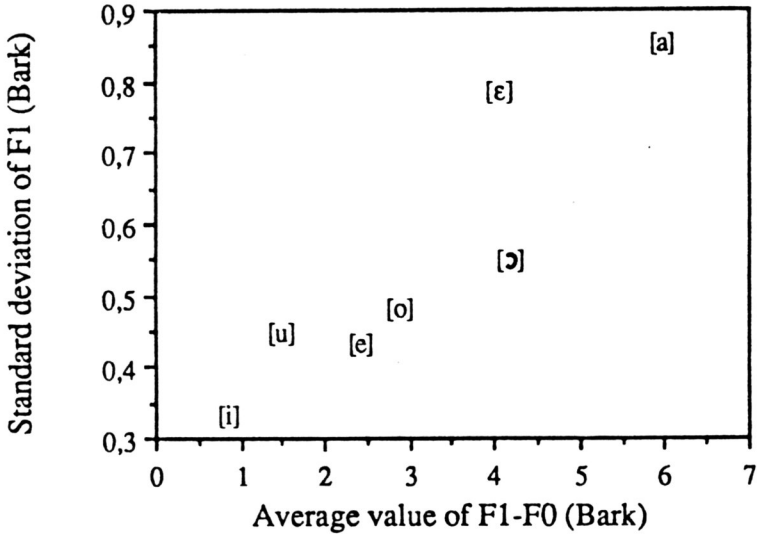

Figure 2 Variability of the first formant F1 as a function of the distance F1-F0

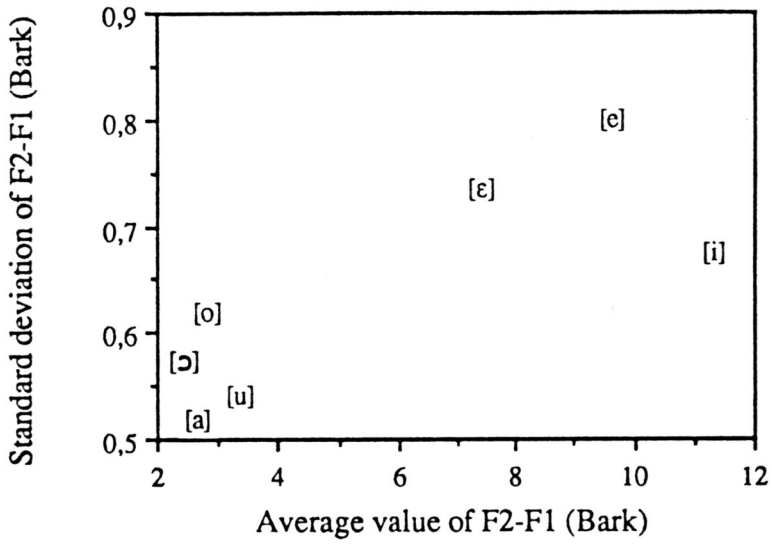

Figure 3 Variability of the distance F2-F1 as a function of the average value F2-F1.

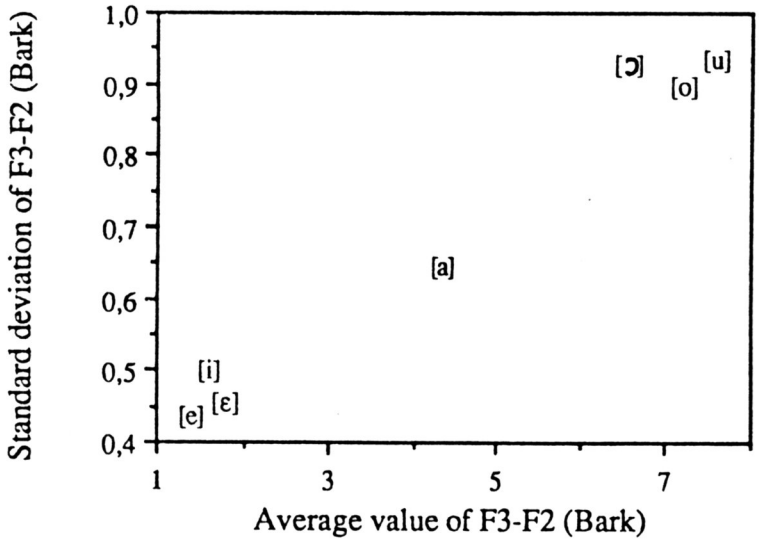

Figure 4 Variability of the distance F3-F2 as a function of the average value of F3-F2.