# EXTRINSIC NORMALIZATION OF VOWEL FORMANT VALUES BASED ON CARDINAL VOWELS MAPPING

Maria-Gabriella Di Benedetto * and Jean-Sylvain Liénard **

* INFOCOM Dept., Facoltà di Ingegneria, University of Rome 'La Sapienza', via Eudossiana 18, 00184 Roma, Italia.

** LIMSI-CNRS, B.P.133, 91403 Orsay cedex, France

ABSTRACT

An extrinsic vowel normalization method is proposed. In the method proposed, a-priori information refering to the first and second formant values (F1 and F2) of three vowels of each speaker is used. In order to obtain the normalized formant values of a given speaker's vowels in a given language it is necessary to:
1.gather reference information on three vowels, for example [i,*a*,u]. This information corresponds to the F1 and F2 values for [i,*a*,u] averaged over a population of native speakers.
2. map the formant values of [i,*a*,u] of the new speaker, onto the reference formant values. This operation can be achieved by linear transformation (isomorphic transformation).
3. the normalized formant values of the other vowels can be obtained by applying the isomorphic transformation.
Results are presented for vowels of American English, and briefly reported for vowels of Italian, and French. Comparison between a classical F1-F2 and the mapped vowel representations in terms of statistical parameters are discussed.

## I. INTRODUCTION

Vowel normalization can be based on either no a-priori knowledge on speaker's characteristics (intrinsic vowel normalization), or on the use of some a-priori knowledge on each speaker's vowel system (extrinsic vowel normalization). In the first category of methods, the perception of a vowel is supposed to be dependent upon the signal itself only, and, in particular, upon few parameters characterizing it. In particular, the formant frequencies and the fundamental frequency are crucial factors in identifying a vowel [1, 2]. The normalization consists in finding a function of these parameters which proves to be invariant with respect to the speaker and to the phonetic context. Several functions, having a normalization effect, have been proposed; most often, the formant and fundamental frequencies are expressed in Mels, Barks, or logarithmic units, and formants ratios or formant differences are considered [3]. In the second category of methods, the perception of a vowel is supposed to be largely influenced by the context in which it is included, implying the existence of an adaptation time to the specific speaker [4, 5].

In the present paper, a method which falls in the second category will be described. In particular, a-priori information refering to the first and second formant values ($F_1$ and $F_2$) of three vowels of a given speaker (for example [i,*a*,u]) is used in order to normalize speaker's vowels.

The method proposed will be described in the second section. This method is based on a 'vowel mapping' transformation of the vowels of a specific speaker onto an 'average' vowel space of a given language. The results obtained by applying this method to vowels of American-English (Peterson and Barney data-base [6]) will be presented in section III. These results will be discussed on the basis of statistical parameters (Mahalanobis distances and linear discriminant analysis classification rates) and compared to those obtained with classical F1 vs F2 representation. In addition, a comparison with the data obtainable using a procedure derived from the method proposed by Gerstman [4] will be discussed. Similar investigations for vowels of Italian and French will be briefly reported. A discussion will be the object of section IV.

this work was supported by a grant of the National Research Councils of Italy (CNR) and France (CNRS).

## II. DESCRIPTION OF THE NORMALIZATION PROCEDURE

The method proposed in the present paper is based on a criterion which can be generally applied in order to map a first set of points onto a second set of points. If the points are defined in a two-dimensional space, and if the coordinates of three reference points are known, it is relatively simple to find the parameters characterizing the mapping transformation.

This criterion can be applied to vowel normalization. The idea is that each vowel of a given speaker can be characterized by its degree of proximity to three vowels (for example [i,*a*,u]) of the same speaker. The proposed method consists in stating the following three equations:

$$\alpha F_1^{[a]} + \beta F_1^{[i]} + \gamma F_1^{[u]} = F_1^{(v)}$$
$$\alpha F_2^{[a]} + \beta F_2^{[i]} + \gamma F_2^{[u]} = F_2^{(v)}$$
$$\alpha + \beta + \gamma = 1$$

where $F_1$ and $F_2$ are the F-patterns of the three reference vowels, and $F_1^{(v)}$ and $F_2^{(v)}$ are the coordinates of the vowel (v) with respect to the three reference vowels, and finding $\alpha$, $\beta$, and $\gamma$.

In matrix form, the above set of equations can be written as follows:

$$\begin{bmatrix} F_1^{[a]} & F_1^{[i]} & F_1^{[u]} \\ F_2^{[a]} & F_2^{[i]} & F_2^{[u]} \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} F_1^{(v)} \\ F_2^{(v)} \\ 1 \end{bmatrix}$$

and equivalently:

$$[A]\,[\theta] = [V]$$

It is therefore possible to compute the vector of coefficients $[\theta]$ for a given vowel by applying the following rule:

$$[\theta] = [A]^{-1}\,[V]$$

If the reference triangle $T_0$ corresponds to the set of $F_1$ and $F_2$ of the three vowels [i,*a*,u], obtained by averaging the formant values of several speakers, it is possible to find the coordinates of the point $V_0$ which will be located with respect to $T_0$ as was V with respect to T. In fact, one has:

$$[V_0] = [A_0]\,[\theta]$$

and therefore:

$$[V_0] = [A_0]\,[A]^{-1}\,[V] = [S]\,[V]$$

Knowing the matrix $[S] = [A_0]\,[A]^{-1}$ it is possible to compute the coordinates of the new point $V_0$. $V_0$ represents the normalized speaker's vowel while V is the unnormalized speaker's vowel.

The matrix [S] represents the isomorphic transformation applied. Figure 1 shows, in a schematic way, the effect of the application of the isomorphic transformation.

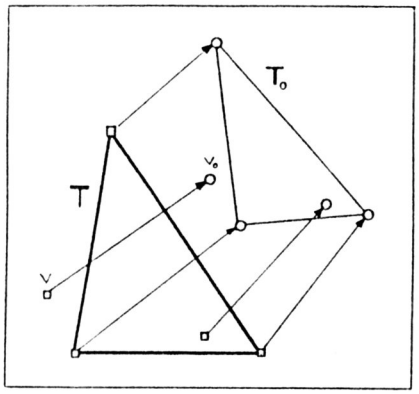


Figure 1. Representation of the isomorphic transformation. The T triangle is transformed into the $T_0$ triangle. The point $V_0$ is located with respect to $T_0$ as was V with respect to T.

## III. EXPERIMENTATION

In this section, results obtained from the application of the method described in the preceding section, are presented. This method was applied to vowels of American-English, Italian, and French. The present paper focuses on the results obtained for the American-English vowels. The results for the other languages will be also briefly reported.

### 3.1 Acoustic data

The acoustic data correspond to the spectrographic measurements made by Peterson and Barney [6] of ten vowels of American English. These vowels were included in the hVd words (heed, hid, head, had, heard, hod, hud, hawed, hood, who'd) and pronounced twice, in isolation, by 28 women, 33 men, and 15 children, leading to 1520 vowel tokens.

### 3.2 Results of the analysis

In this paragraph, results of the analysis on the data described in the previous paragraph will be presented. The vowel areas will be represented in terms of elliptic areas in a two-dimensional space. This type of representation is valid under the assumption that the parameters corresponding to the two dimensions ($P_1$ and $P_2$) have a gaussian probability density function (pdf). Under this assumption, if one considers the intersection of the gaussian pdf of $P_1$ and $P_2$ and a plane parallel to the $P_1$-$P_2$ plane, the projection of this intersection on the $P_1$-$P_2$ plane is an ellipse. These ellipses can be thought as loci of equiprobability P; P corresponds to the probablity of being "inside the ellipse", while 1-P is the probability of being "outside the ellipse".

The data will be presented in terms of ellipses of equiprobability with P=0.7. Figure 2 shows the acoustic data in the $F_1$ vs $F_2$ plane for the vowels [i,*a*,u] and the ellipses of equiprobability corresponding to P=0.7.

The results of the analysis in the $F_1$ vs $F_2$ plane for all vowels of the Peterson and Barney data-base are presented in Fig.3. Figure 3 shows that ('statistically') most of the contiguous vowel areas overlap. In order to quantify the amount of overlap, the Mahalanobis distance between vowel areas in the $F_1$ vs $F_2$ plane was computed. This distance indicates the actual euclidean distance between the mean vectors of the two sets, normalized with respect to the covariance matrices of the two sets. The Mahalanobis distances between vowel areas in the F1 vs F2 plane, are presented in Table I. Note that when two vowels ellipses do not overlap, the Mahalanobis distance is usually greater than 10.

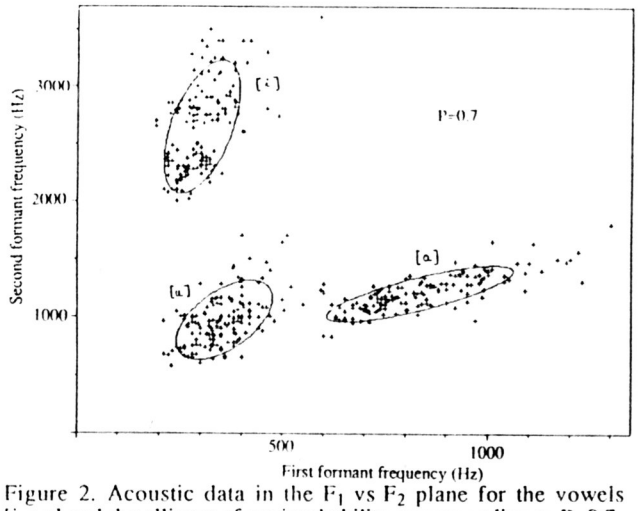


Figure 2. Acoustic data in the $F_1$ vs $F_2$ plane for the vowels [i,*a*,u] and the ellipses of equiprobability corresponding to P=0.7.

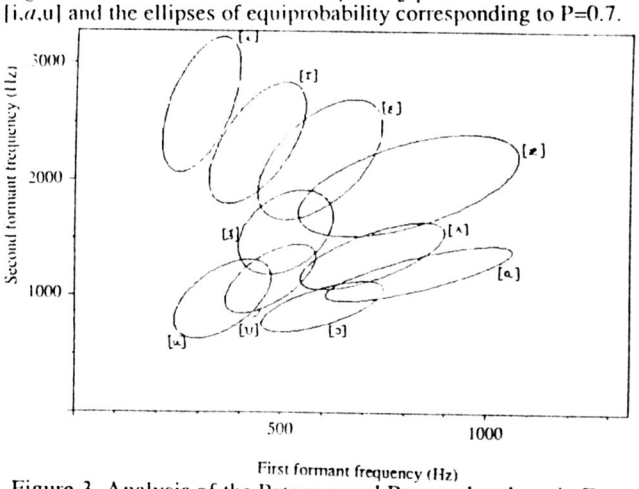


Figure 3. Analysis of the Peterson and Barney data-base in $F_1$ vs $F_2$ plane (ellipses of equiprobability corresponding to P=0.7).

Table I - Mahalanobis distances between vowels in F1 vs F2 plane.

| | [I] | [ɛ] | [æ] | [*a*] | [ʌ] | [o] | [U] | [u] | [ɜ] |
|---|---|---|---|---|---|---|---|---|---|
| [i] | 10 | 27 | 34 | 82 | 92 | 95 | 65 | 47 | 39 |
| [I] | | 6 | 17 | 50 | 66 | 64 | 32 | 26 | 14 |
| [ɛ] | | | 6 | 23 | 38 | 36 | 15 | 18 | 5 |
| [æ] | | | | 8 | 19 | 24 | 12 | 17 | 5 |
| [*a*] | | | | | 6 | 8 | 8 | 15 | 11 |
| [ʌ] | | | | | | 4 | 16 | 20 | 23 |
| [o] | | | | | | | 12 | 14 | 20 |
| [U] | | | | | | | | 2 | 4 |
| [u] | | | | | | | | | 7 |

The same data-base was transformed according to the procedure proposed in the present paper. Results are presented in Fig.4. Figure 4 shows the vowel ellipses in the normalized $F_1$-$F_2$ plane. The three vowels used to apply the normalization were [i,*a*,u]. Other triplets of vowels could have been selected. However, although a detailed description on the choice of the reference vowels is beyond the scope of the present paper, note that it is preferable to choose a vowel triplet which covers as much as possible the formant space. Note on Fig.4 that each reference vowel was, after normalization, represented by the same point. It appears after Fig.4 that, globally, the normalization helped reducing the spread of the vowel areas. In order to quantify this difference, the Mahalanobis distances between vowels in the normalized F1-F2 plane were computed, and the values obtained were compared to the results of Table I. These distances are presented in Table II. Table II does not include the distances refering to [i,*a*,u] since these vowels were the very basis of the normalization procedure.

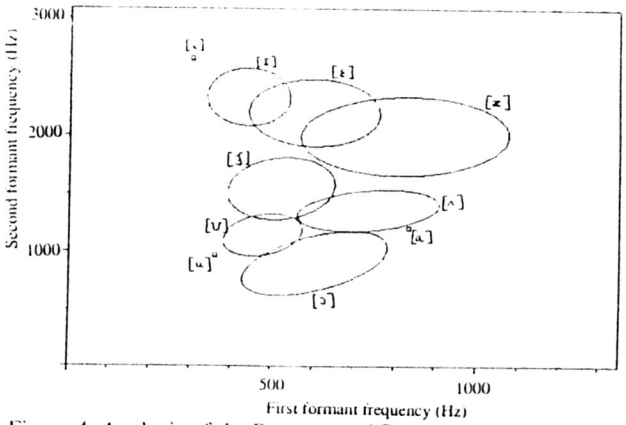


Figure 4. Analysis of the Peterson and Barney data-base in $F_1$ vs $F_2$ normalized plane (ellipses of equiprobability corresponding to P=0.7).

Table II - Mahalanobis distances between vowels in F1-normalized vs F2-normalized plane. The normalization was obtained with respect to the three vowels [i,*a*,u]. Consequently, the distances for these three vowels are not significative and are not reported.

| | [I] | [ɛ] | [æ] | [ʌ] | [*a*] | [o] | [U] | [u] | [ɜ] |
|---|---|---|---|---|---|---|---|---|---|
| [i] | * | * | * | * | * | * | * | * | * |
| [I] | | 4 | 13 | 67 | * | 95 | 75 | * | 25 |
| [ɛ] | | | 4 | 33 | * | 54 | 47 | * | 14 |
| [æ] | | | | 13 | * | 30 | 30 | * | 10 |
| [ʌ] | | | | | * | 9 | 9 | * | 8 |
| [*a*] | | | | | | * | * | * | * |
| [o] | | | | | | | 8 | * | 18 |
| [U] | | | | | | | | * | 7 |
| [u] | | | | | | | | | * |

The results obtained using the method proposed in the present paper were compared to those obtainable by applying a procedure derived from Gerstman's method [4]. The method proposed by Gerstman was based on the normalization of the $F_1$ and $F_2$ values of all vowels with respect to the maximum and minimum $F_1$ and $F_2$ values among all vowels. Consequently, this method made use of a-priori information on all vowels of the vowel system. Since the method proposed in the present paper makes use of a-priori information on three vowels only, a modification of the Gerstman's method was derived (MG procedure). In this modified procedure, the vowel formants were normalized with respect to the $F_1$ and $F_2$ maximum and minimum values among the three reference vowels [i,*a*,u] only. Figure 5 shows the results obtained using this method. The Mahalanobis distances in this case are reported in Table III.

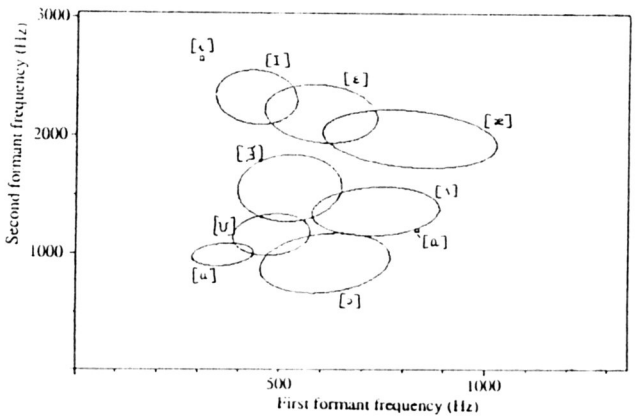


Figure 5. Analysis of the Peterson and Barney data-base in $F_1$ vs $F_2$ MG-normalized plane (ellipses of equiprobability corresponding to P=0.7).

Table III - Mahalanobis distances between vowels in F1-MG-normalized vs F2-MG-normalized plane. The normalization was obtained with respect to the three vowels [i,*a*,u]. Consequently, the distances for these three vowels are not significative and are not reported.

| | [I] | [ɛ] | [æ] | [ʌ] | [*a*] | [o] | [U] | [u] | [ɜ] |
|---|---|---|---|---|---|---|---|---|---|
| [i] | * | * | * | * | * | * | * | * | * |
| [I] | | 5 | 15 | 60 | * | 88 | 78 | * | 23 |
| [ɛ] | | | 5 | 33 | * | 61 | 58 | * | 14 |
| [æ] | | | | 19 | * | 46 | 50 | * | 14 |
| [ʌ] | | | | | * | 9 | 11 | * | 7 |
| [*a*] | | | | | | * | * | * | * |
| [o] | | | | | | | 6 | * | 15 |
| [U] | | | | | | | | * | 7 |
| [u] | | | | | | | | | * |

Comparing the results reported in Tables I, II, and III, it is evident that both normalization methods helped reducing the spread of the vowels under consideration. However, comparing the results obtained after normalization (Tables II and III), does not lead to a clear superiority of one over the other. The same conclusion can be drawn after examination of the vowel ellipses (Figs. 3, 4, and 5).

In order to better evaluate the differences between the two normalization methods, a classification procedure was applied to the vowel data. This procedure will be described in the next paragraph.

3.3 Vowel classification

A linear discriminant analysis was carried out in order to quantify the differences between the representations described in the previous paragraph. This analysis consists in finding the linear function which best separates two sets. When the sets are defined in a two-dimensional space, this linear function is a straight line. Figure 6 shows an example of linear discriminant function between two groups.

Once the linear discriminant function has been determined, it is possible to classify a vowel token as belonging to one of the two sets. The results of the linear discriminant analysis can thus be given in terms of misclassification rates.

The linear discriminant analysis was carried out on the three measurement sets already described in the previous paragraph: F1-F2 measurements (refering to the Peterson and Barney data), F1 normalized-F2 normalized measurements (refering to the Peterson and Barney data after normalization according to the method proposed in the present paper), F1-MG-normalized and F2-MG-normalized data ( refering to the Peterson and Barney data after normalization according to the modification of Gerstman's method).

The results of this analysis are given in terms of misclassification scores. Table IV shows the results in the case of the F1-F2 measurements, Table V shows the results in the case of the F1 normalized-F2 normalized measurements, and Table VI shows the results in the case of F1-MG-normalized and F2-MG-normalized data.

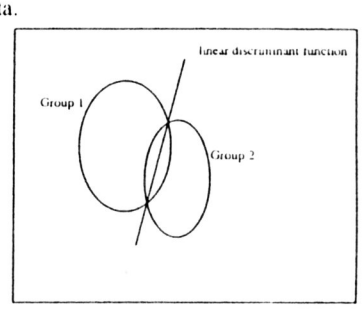


Figure 6. An example of a linear discriminant function.

Examination of Tables V and VI indicates a clear superiority of the first normalization method. Averaging the misclassification scores over all vowel pairs led to: 4.9% for the unnormalized values, 2.8% for the normalization method proposed in the present paper, and 3.4% for the modified

Table IV - Misclassification scores in the case of the F1 and F2 values.

| | [I] | [ɛ] | [æ] | [ɑ] | [ʌ] | [o] | [U] | [u] | [ɜ] |
|---|---|---|---|---|---|---|---|---|---|
| [i] | 6.6 | .3 | .3 | .0 | .0 | .0 | .0 | .0 | .3 |
| [I] | | 11.2 | 2.0 | .0 | .0 | .0 | .0 | .3 | 2.3 |
| [ɛ] | | | 9.2 | .0 | .0 | .0 | 1.0 | 1.0 | 16.4 |
| [æ] | | | | 3.9 | .0 | .0 | 1.6 | 1.0 | 14.5 |
| [ɑ] | | | | | 8.6 | 8.2 | 6.9 | 1.6 | 2.6 |
| [ʌ] | | | | | | 14.8 | .3 | .7 | .3 |
| [o] | | | | | | | 3.3 | 2.6 | .3 |
| [U] | | | | | | | | 17.1 | 19.1 |
| [u] | | | | | | | | | 8.9 |

Gerstman's method. Note that each vowel set was composed by 152 tokens and the discriminant analysis is applied to 21 different vowel pairs, leading to 6384 classifications.

Table V - Misclassification scores between contiguous vowels in F1-normalized vs F2-normalized plane. The normalization was obtained with respect to the three vowels [i,ɑ,u]. Consequently, the misclassifications for these three vowels are not significative and are not reported.

| | [I] | [ɛ] | [æ] | [ʌ] | [ɑ] | [o] | [U] | [u] | [ɜ] |
|---|---|---|---|---|---|---|---|---|---|
| [i] | * | * | * | * | * | * | * | * | * |
| [I] | | 13.2 | 2.6 | .0 | * | .0 | .0 | * | 1.6 |
| [ɛ] | | | 12.2 | .3 | * | .0 | .0 | * | 1.6 |
| [æ] | | | | 1.3 | * | .0 | .0 | * | 1.6 |
| [ʌ] | | | | | * | 1.6 | 3.0 | * | 3.9 |
| [ɑ] | | | | | | * | * | * | * |
| [o] | | | | | | | 6.9 | * | .7 |
| [U] | | | | | | | | * | 9.2 |
| [u] | | | | | | | | | * |

Table VI - Misclassification scores between contiguous vowels in F1-Gerstman-normalized vs F2-Gerstman-normalized plane. The normalization was obtained with respect to the three vowels [i,ɑ,u]. Consequently, the scores for these three vowels were not significative and were not reported.

| | [I] | [ɛ] | [æ] | [ʌ] | [ɑ] | [o] | [U] | [u] | [ɜ] |
|---|---|---|---|---|---|---|---|---|---|
| [i] | * | * | * | * | * | * | * | * | * |
| [I] | | 13.2 | 3.0 | .3 | * | .0 | .0 | * | 1.6 |
| [ɛ] | | | 11.8 | .7 | * | .0 | .7 | * | 3.0 |
| [æ] | | | | 2.0 | * | .0 | .0 | * | 2.6 |
| [ʌ] | | | | | * | 6.6 | 3.0 | * | 4.9 |
| [ɑ] | | | | | | * | * | * | * |
| [o] | | | | | | | 8.2 | * | 1.3 |
| [U] | | | | | | | | * | 9.2 |
| [u] | | | | | | | | | * |

### 3.4 Results for other languages

The above methods were also applied to French and Italian vowel data. As regards the French case, the data-base consisted of 9 oral vowels [i,e,ɛ,y,ø,œ,a,o,u] pronounced in isolation with three different vocal efforts by 5 female and 6 male speakers. The speech tokens were uttered in a relatively natural manner since in French all the selected vowels form lexical words [7]. Results showed that both transformations yielded a reduction in the spread of the vowel areas. The vowel areas were also made closer and consequently the discrimination between vowels remained similar.

As regards the Italian case, the data-base consisted of 7 vowels [i,e,ɛ,a,o,o,u] pronounced in isolation by 25 male and 11 female speakers [8]. Isolated vowels form meaningful words in Italian. According to the instructions received, the vowels were pronounced in a sustained way. In this case, the reduction in vowel areas spread was large with both transformations. The number of misclassifications was also reduced but by a smaller amount. The results showed a slight advantage for the isomorphic transformation over the MG method.

## IV. DISCUSSION

The underlying hypothesis of the present study was that the whole vowel system of a given speaker in a given language could be derived from the knowledge of three vowels only. A method was proposed which allowed the prediction of the $F_1$-$F_2$ values for a new vowel, knowing the $F_1$-$F_2$ values for this vowel in the reference system. If this hypothesis were valid, it would have consequences in several fields: in language training, to help a student to produce the right sound [9]; in automatic speech recognition, to contribute to speaker adaptation; in speech synthesis, to produce different sounding-voices.

The results of the analysis showed that the hypothesis was partially verified. On the one side, both transformations produced a significant reduction of the overlapping between adjacent vowel areas; in particular the variations due to speaker gender (male, female, children) were largely compensated for. On the other side, a great variability within each vowel category still persisted. This variability could be attributed to different factors such as differences in dialects, pronunciation habits, speaking style. The understanding of whether these causes could be identified and restricted to a small number, or whether multiple unidentifiable factors contributed to the observed variability, remains unclear and needs further investigation.

As for the differences between the two methods, note that on the basis of the vowel ellipses representation and of the Mahalanobis distances, the two methods were quite similar. In fact, these analysis tools were based on the hypothesis that the pdf of the measurements considered be Gaussian. Consequently, the models corresponding to the vowel data, more than the vowel data itself, were compared. In the linear discriminant analysis, the statistical modeling played a minor role: the vowel tokens were compared and classified individually.

Both methods attempt to map different vowel sets in order to reduce the variability of the measurements within each set. However, the MG method produces translation and rescaling of the vowel areas while the isomorphic transformation also produces a rotation combined with the two above effects. Consequently, after transformation a triangle could change its proportions, although the topology of the system was kept. In addition the MG method is simpler and more robust with respect to the $F_1$-$F_2$ variations than the isomorphic transformation. The two methods also differ in nature since the MG method applies a non linear criterion (choice in $F_1$ and $F_2$ limits) while the isomorphic transformation is linear.

Beyond the geometrical considerations, the main difference between the two methods is that the MG method is based on the independency of $F_1$ and $F_2$. The isomorphic transformation considers the relations between the vowels as forming a whole system according to some rules imposed by the structure of the language.

## REFERENCES


[1] R. R. Verbrugge, W. Strange, D. P. Shankweiler, and T. R. Edman. "What information enables a listener to map a talker's vowel space?", *J.Ac.Soc.Am.* 60(1), pp.198-212, 1976.

[2] M. J. Macchi. "Identification of vowels spoken in isolation vs vowels spoken in consonantal context", *J.Ac.Soc.Am.* 68(6), pp. 1636-1642, 1980.

[3] A.K. Syrdal and H. S. Gopal. "A perceptual model of vowel recognition based on the auditory representation of American English vowels", *J.Ac.Soc.Am.* 79(4), pp.1086-1100, 1986.

[4] L. H. Gerstman. "Classification of self-normalized vowels", *IEEE Trans. Audio Electroac.* AU-16, pp.78-80, 1968.

[5] J. Hillenbrand and R.T. Gayvert. "Speaker-independent vowel classification based on fundamental frequency and formant frequency", 113th ASA meeting, Indianapolis, 1987.

[6] G.E. Peterson and H.L. Barney. "Control methods used in a study of the vowels",*J.Ac.Soc.Am.* 24(2), pp.175-184,1952.

[7] J.S. Liénard and M.G. Di Benedetto. "Evaluation perceptive d'un corpus de voyelles françaises émises isolément par plusieurs locuteurs selon diverses forces de voix", 19e J.E.P., Bruxelles, pp.469-474, 1992.

[8] M.G Di Benedetto and G. Flammia. "Vowel distinction along auditory dimensions: a comparison between a statistical and a neural classifier", Verba 90, Rome, 1990.

[9] M.G. Di Benedetto, F. Carraro, S. Hiller, and E. Rooney. "Vowels pronunciation assessment in the SPELL system", ICSLP, Banff, Alberta, Canada, 1992.