CHARACTERIZATION OF THE NON-LINGUISTIC INFORMATION OF VOWELS BY MATCHING VOWEL SYSTEMS

Jean-Sylvain Liénard, LIMSI-CNRS, Orsay, France
Maria-Gabriella Di Benedetto, INFOCOM, Univ. La Sapienza, Rome, Italy

ABSTRACT

Vowel system normalization does not succeed, in general, in totally cancelling the scattering areas of vowels. Considerable variations remain, which are due to the speaker and context peculiarities. In the present study we consider all types of information as equally important in the analysis of speech structures. Using the Peterson and Barney data we show that, by matching vowel systems, linguistic information can be associated with an average system considered as reference, while non-linguistic information lies partly in the parameters of the transform which gets an individual system close to the reference, and partly in the deviation of the transformed system with respect to the reference.

GENERAL PRESENTATION

Previous work on Vowel Systems (VS) normalization from three cardinal vowels resulted in some reduction of the non-cardinal vowels scattering areas, but failed in eliminating all classification errors [1]. The very notion of normalization implies some limitations : imposing too close a match between two VSs yields to neglect some relevant discrepancies among languages, dialects or individuals [2]. We propose here to compare VSs to each other as wholes, and to look into the parameters of the matching transformation for what corresponds to linguistic information (i.e. the explicit phonetic code of the language) as well as for what corresponds to non-linguistic information (i.e. the identity and vocal gender of the talker, the type of voice, etc.). This approach is an illustration of the Speech Pattern Processing paradigm [3], according to which all aspects of the perceptive information of the signal must be taken into account simultaneously.

From Peterson and Barney's vowel formant measurements [4] we determine a Reference Vowel System (RVS) ; then we define several transforms aiming at a "best match" of the RVS with the individual VSs. For each transform we compute the error-rate obtained in the classification of the vowels, speakers, and vocal gender (male, female, child). Finally we discuss the ability of each transform to provide an adequate representation of the relevant information.

CORPUS AND TRANSFORMS

Peterson and Barney's data comprise 10 American vowels uttered twice by 76 speakers (33 males, 28 females and 15 children). We only use the F1 and F2 measurements. The RVS is arbitrarily obtained by averaging the male VSs, which are the most represented in the database, and the less subject to formant frequency measurement errors.

We look for a transform which achieves an optimal matching of two sets of homologous points RVS and VS in the (F1,F2) coordinates (fig 1). In order to get some generality this transform must be simple : it is made of scalings and translation. Thus both systems cannot be exactly mapped onto each other : the direct transform changes RVS into a new system which best approximates VS ; the inverse transform changes VS into a new system, called Inverse Vowel System (IVS), which best approximates RVS. The quality of the approximation refers to the mean quadratic distance between homologous points of both systems,

measured in the (F1,F2) plane, with a weight of 2 in favor of the F1 dimension, to compensate for the interval (in Hertz) between extreme values, which is about half for F1 than it is for F2. The translation which gets a system as close as possible to the other is defined by the vector joining both centers of mass.

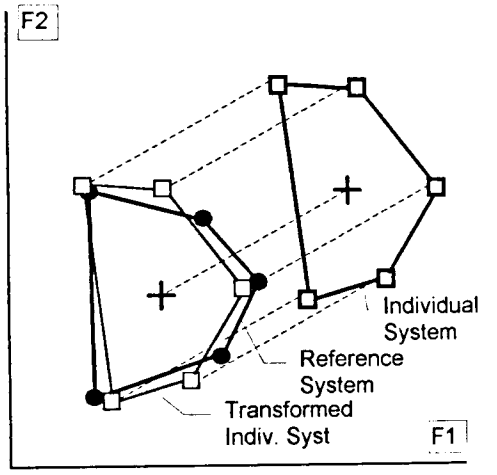

fig 1 : Vowel systems and transform

Such a translation, done with F1 and F2 replaced by their logarithms, corresponds to the multiplication of all formant frequencies by the same factor. It looks quite natural, in view of the perceptual theories based on the constancy of the formant frequencies ratio for a given vowel (cf, among other authors, Nearey's log-mean normalization reported in [2] and the formant-ratio theory evoked by Miller [5]). This transform, termed "Simple Log", is defined for each VS by the value of a single parameter (log of the best scaling factor in F1 and F2).

The "Double Log" transform does not presuppose this constancy : it applies a translation according to the F1 dimension, and another one in the F2 dimension. Thus it is defined by two parameters (logs of the best scaling factors in F1 and in F2).

Two other transforms, "Simple Bark" and "Double Bark", are similarly defined by replacing the log by the Bark function : close to the linear scale in the low frequency range and close to the log scale in the high frequency range.

We also used the Gerstman transform, which consists of linearly mapping the interval between extreme values in the F1 dimension of both systems ; the same process holds independently in the other dimension [6]. This transform requires 4 parameters.

The case for which the original data is used, i.e. no transform is applied, is called the Null transform.

For a given transform we make the following calculations. Each VS of the database transformed into an IVS as close as possible to the RVS. The parameters are kept, as well as the set of deviations (in both coordinates) that remain between homologous points of RVS and IVS. Thus there is no loss of information : each VS can be exactly reconstructed knowing the RVS, the parameters and the set of deviations. Then several error-rates are computed, concerning the vowel identity, vocal gender and talker identity (table 1).

table 1: error-rate in the classification according to several kinds of information and several transforms

transform	vowel err. from distances	gender.err. from params	talker err. from params	talker err. from dev. sets
random	*90.0*	*63.7*	*99.3*	*99.3*
Null	33.8	-	-	17.8
Simple Log	13.5	8.6	86.8	28.9
Double Log	11.5	9.2	65.8	36.8
Simple Bark	14.6	7.9	80.3	27.6
Double Bark	11.2	7.9	69.7	37.5
Gerstman	16.6	9.9	73.0	50.7

RESULTS

Errors on vowel categories

Each token of the IVS is compared to the RVS tokens (computation is based on the distances after transformation). The closest one is selected. If the vowel labels do not match an error is counted. Computation is extended to the whole database.

All the transforms improve the initial situation, which is normal. Double Log and Double Bark, with two parameters each, perform best. Simple Log and Simple Bark yield slightly degraded results but remain surprisingly efficient given the fact that they use one single parameter. Finally Gerstman seems less efficient despite its 4 parameters. This can be attributed to the fact that only 3 or 4 vowels, out of 10, contribute to the determination of the transform, making it non optimal for the whole VS. The same remark holds true in the further experiments.

Errors on the vocal gender computed from the transform parameters

First three average values of the parameters are computed, each one relating to a vocal gender, on the whole database. Then for each VS a quadratic distance is computed (on the parameters) to the three gender representatives, and the closest one is selected. An error is counted when the gender labels do not match. All transforms practically give the same result (8 to 10%), which seems good as compared to random gender allocation (63.7%). The most interesting observation is that one parameter is sufficient to characterize this aspect ; adding other parameters does not change anything.

Errors on talker identification from the parameters

For each VS the parameter values are compared using a quadratic distance to the parameter values of the 151 other systems of the database. The closest one is selected. If the talker labels do not match an error is counted.

Globally the error-rates are high, which shows that the parameters do not capture much talker-specific information. Double Log and Double Bark perform slightly better than their single-parameter conterparts.

Errors on talker identification from the deviation sets

For each VS the deviation set (10 components in F1, 10 in F2) is compared using a quadratic distance to the deviation sets of the 151 other systems of the database. The closest one is selected. If the talker labels do not match an error is counted.

Contrary to the previous experiment it appears that the deviation set captures most of the talker-specific information. Compared to the others, the one-parameter transforms leave more room to expressing the individual talker peculiarities in the proper shape of the transformed systems and consequently in the deviation sets.

On the relative merits of the transforms

Leaving aside the Gerstman transform for the above-mentioned reason, our experiments show that

- most of the gender information lies in a single parameter, namely the log-mean or bark-mean ratio,
- using two parameters instead of one results in assigning more talker-specific information to the parameters, more vowel-specific information and less talker-specific information to the transformed systems,
- the log and bark scales yield comparable results.

CONCLUSION

By considering the matching of vowel systems as wholes we showed, using the Peterson and Barney data, that linguistic and diagnostic kinds of information could be - at least partially - decorrelated.

Linguistic information, common to a group of talkers, lies mostly in the relative disposition of the vowels after some talker- or gender-specific transformation. This disposition may be materialized in the Reference Vowel System by vocalic categories defined by their prototypes. When a particular Vowel System is transformed so that it best approximates the Reference Vowel System the vowel error-rate drops below 15%.

Diagnostic information, related to the talker or voice specification, can be found partly in the transform parameters, partly in the set of deviations of the transformed system with respect to the Reference Vowel System. The proportions vary according to the number of parameters, that is the ability of the transform to match both systems more closely. It is remarquable that a single parameter, namely the scaling factor or its logarithm, conveys most of the information about the vocal gender.

REFERENCES

[1] Di Benedetto, M.G. and Liénard, J.S., "Extrinsic normalization of vowel formant values based on cardinal vowels mapping", ICSLP, Banff, Oct 12-16, 1992.

[2] Ferrari-Disner, S., "Evaluation of vowel normalization procedures", J.Acoust.Soc.Am. 67(1), 253-261, 1980.

[3] Liénard, J.S., "Speech Pattern Processing : integrating the linguistic and non-linguistic aspects of voice and speech", ICPhS, Stockholm, Aug. 13-19, 1995.

[4] Peterson, G. and Barney, H., "Control methods used in a study of the vowels", J.Acoust.Soc.Am. 24, 175-184, 1952.

[5] Miller, J.D., "Auditory-perceptual interpretation of the vowel", J.Acoust. Soc.Am. 85(5), 2114-2134, May 1989.

[6] Gerstman, L.J., "Classification of self-normalized vowels", IEEE Trans. on Audio and Electroac., AU-16, 1, 78-80, 1968.

Research supported by a cooperation agreement between the French Centre National de la Recherche Scientifique and the Italian Consiglio Nazionale delle Ricerche.