VOWEL CLASSIFICATION BASED ON ACOUSTIC AND ARTICULATORY REPRESENTATIONS

Alain Soquet[1] *and Marco Saerens*[2]
[1]*Institut des Langues Vivantes et de Phonétique,* [2]*IRIDIA,*
Université Libre de Bruxelles

ABSTRACT

The objective of this paper is to compare different acoustic and articulatory representations on a vowel classification task. Classification results were obtained based on linear discriminant analysis and decision trees algorithm with cross-validation on the speakers. The cepstrum, the formants and the articulatory representations achieve similar performances with linear discriminant analysis. The decision tree algorithm provides accurate classification rules for the formants and the articulatory representations. The resulting articulatory rules are consistent with our knowledge on vowel production and could be efficiently used in knowledge-based systems.

INTRODUCTION

A problem of long-standing interest in speech analysis and recognition concerns the most appropriate representation for acoustic-phonetic decoding. In this work we will focus on vowels; results for plosives place of articulation identification can be found in [1]. Vowels are traditionally described in term of static spectral characteristics or articulatory configurations.

From the acoustic point of view, global spectrum descriptions in term of a small set of coefficients (for example the LPC coefficients) can be used. However, the standard representation for vowels consists in the first resonance frequencies of the vocal tract (the formants frequencies). It is well known that, even if the first formants frequencies are efficient cues to classify vowels, there exist an important speaker variability – for example, the differences between male and female speakers [2]. Moreover, in fluent speech, the target vowels are not always reached when produced in a consonant context [3]. Nevertheless, this phenomenon does not introduce any degradation in the human recognition capabilities [4]. A large amount of work continues however to focus on static description of vowels (see for example [5] and [6]).

From the articulatory point of view, one can describe a vowel by the configuration of an articulatory model that produces similar spectral characteristics. Unfortunately, the computation of the articulatory configuration from the acoustic parameters (the acoustic-to-articulatory inversion) is not a trivial problem. In previous work [7], we developed a tool that realizes this inversion in the framework of an articulatory model, based on the first three formant frequencies.

Our aim in this work is to compare several acoustic and articulatory representations on a vowel classification task.

ACOUSTIC REPRESENTATIONS

The speech signal was passed through a 5 kHz cutoff low-pass filter, and sampled at 10 kHz. The signal was then preemphasized (1 - 0,95 z^{-1}) before further processing. Six different acoustic representations have been chosen. Three of them are directly computed from the speech signal, and are widely used in statistical speech recognition systems (e.g. HMM). The three remaining ones are related to formant frequencies, prevalent in knowledge-based recognition systems.

- **LPC** (LPCA): The LPC coefficients were computed with the autocorrelation method on a 25,6 ms frame multiplied by a Hamming window. The number of poles of the predictive filter was fixed to 12.
- **LPC cepstrum** (LCPS): The LPC cepstral coefficients were derived from the predictive coefficients obtained with an LPC analysis [8]. As before, we used the first 12 coefficients.
- **Cepstrum** (CPST): Cepstral coefficients were computed from a 16 ms frame multiplied by a Hamming window. The first 12 coefficients of the cepstrum were used in order to describe the spectral characteristics of the signal at the measurement point.
- **Formants** (FORM, BARK, MEL): The formant values were extracted semi-automatically on the basis of the different acoustic representations. We used 3 different scales for the frequency axis: Hertz (FORM), Bark [9] (BARK), and Mel [10] (MEL).

ARTICULATORY REPRESENTATIONS

Four articulatory representations were selected. The first one is computed from the LPC coefficients. The other three correspond to the control parameters of three different articulatory models. These control parameters are provided by a neural network performing the acoustic-to-articulatory inversion on the basis of the first three formant frequencies [7].

- **LPC area** (LAREA): The LPC area functions are computed from the LPC reflection coefficients as suggested by [11].
- **DRM** (DRM): The distinctive regions model [12] is an 8 regions acoustic tube with transversal control. The model is derived from acoustic properties of the uniform acoustic tube. The control parameters are the sections of the 8 regions. The length of the tube was kept constant (18 cm).
- **Maeda** (MAEDA): Maeda's model [13] is an articulatory model derived from X-ray sagittal cuts. A set of 7 parameters controls the shape of the sagittal cuts.
- **Lin Fant** (LF): Lin and Fant's model [14] is a geometrical model with longitudinal control. There are 3 main control parameters (two for the principal constriction, and one for the lips).

EXPERIMENTS

In order to study the effectiveness of these different representations for the classification of vowels, a set of vowel-consonant-vowel (V^1CV^2) was recorded (where C is one of the six plosives [p, t, k, b, d, g], and V^1 or V^2 one of the five vowels [a, œ, i, u, y]). The resulting 150 VCV were recorded by 11 male speakers, giving a total of 1650 tokens and 3300 vowels. The 10 representations are computed in the stable part of the vowel V^1 and V^2.

In a first experiment, vowel recognition results were obtained based on linear discriminant analysis [15] with cross-validation on the speakers: the tokens from each individual speaker are successively removed from the training set, and used as a test set. The results can therefore be considered as speaker-independent. Each training set consists in 3000 vowels, and each test set in 300 vowels. The results are presented in figure 1.

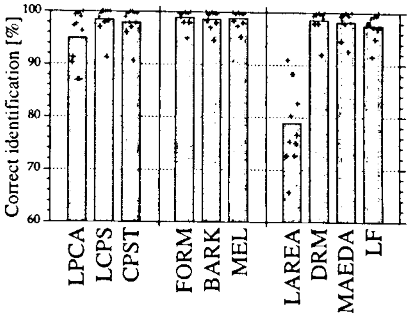

Figure 1: Results of discriminant analysis for the ten representations showing percent of correct classification: results for each test speaker (crosses) and averaged performance (in grey).

The formant based representations obtain the best performances on average. Their scores are, however, comparable with those of other acoustic representations – like LCPS – which have higher dispersions. We can observe that LPCA obtains lower performances.

The performances of LAREA – less than 80 % and a very high variability among the different speakers – are significantly lower than the other representations. On the contrary, the three other articulatory representations present performances comparable to the formant cues. Their dispersions are larger than for FORM but remain lower than for LCPS. On average, LF is less performant than DRM and MAEDA.

We observed that the vowel giving the largest amount of errors is [a], often confused with [œ]. This result is illustrated in figure 2 showing the scatter plot of the 3300 vowels on the two first discriminant axes for 4 representations.

In a second experiment, we used a decision trees algorithm named C4.5 [16]. This technique allows to build a tree that classify the data with a succession of tests involving just one attribute. The tree is

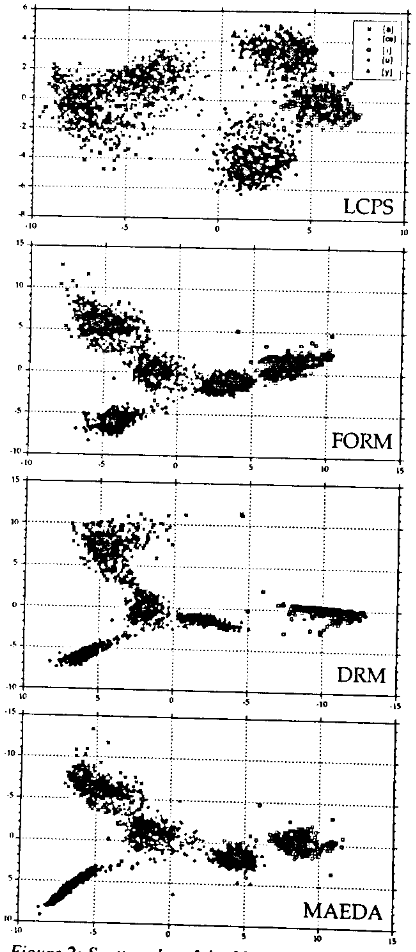

Figure 2: Scatter plot of the 3300 vowels on the two first discriminant axes for LCPS, FORM, DRM and MAEDA.

then pruned so that it becomes both simpler and more accurate on unseen cases. Finally, the algorithm generates a production rule classifier that is usually as accurate as the pruned tree, and more easily understood by people. This algorithm has been applied to four representations: LCPS, FORM, DRM and MAEDA. As for the discriminant analysis, the performances were obtained with cross-validation on the speakers. The results are presented on figure 3.

LCPS obtains bad performances in comparison with the other three representations. The performances of LCPS are even worse when using the rules. This result can be explained by the size of the trees generated by the algorithm (see table 1). The trees for LCPS are on average three times larger than for the other representations and their interpretation is quite intricate. Therefore, the algorithm does not succeed in generating a set of rules able to classify efficiently the five vowels. This indicates that the boundaries are complex and that the vowels cannot be separated with simple production rules.

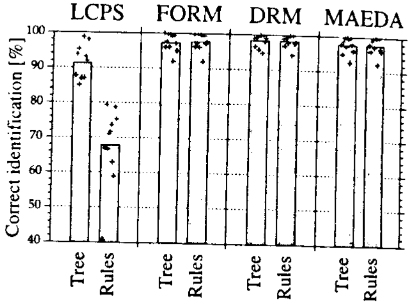

Figure 3: Results of decision trees and rules for four representations showing percent of correct classification: results for each test speaker (crosses) and averaged performance (in grey)

Table 1: Comparison of results obtained with the decision trees (mean size of the trees – # – and correct classification scores) and with the rules after pruning the tree (mean number of rules – # – and correct classification scores).

	Decision trees		Rules	
Method	#	Classified	#	Classified
LCPS	121.5	91.28 %	8.8	67.83 %
FORM	39	97.43 %	9	97.66 %
DRM	25.7	98.25 %	6	98.31 %
MAEDA	34.3	97.67 %	7.3	97.67 %

On the contrary, the performances of the three other representations are similar to those obtained by discriminant analysis. Moreover, the rules generated by the algorithm are quite intuitive :

- The rules deduced from FORM make efficient use of the formant frequencies in order to discriminate the vowels.
- The rules deduced from the two articulatory representations DRM and MAEDA are very intuitive and consistent with our knowledge on the production of vowels. They use the main constriction and the lips opening to distinguish among the vowels.

Finally, it is interesting to note the small size of the tree for DRM, able to classify the five vowels with, on average, only six rules.

CONCLUSIONS

We compared 10 representations of the speech signal on a vowel identification task with two different classification procedures: the linear discriminant analysis and the decision trees algorithm. The cepstrum, the formants and the articulatory representations achieve similar performances with linear discriminant analysis. When using the decision tree algorithm, similar performances are only obtained for formant and articulatory representations. Indeed, for the cepstrum, the performances of the rule-based classifier are found to be significantly worse. This can be explained by an overfitting of the training set which results in very complex trees that are unable to abstract the data.

ACKNOWLEDGMENTS

This work was partially supported by the "Communauté Française de Belgique" and the "European Communities" in the framework of the ARC 93/98 — 168, ARC 92/97 — 160, and FALCON (6017) Basic research ESPRIT projects.

REFERENCES

[1] A. Soquet, and M. Saerens, "A comparison of different acoustic and articulatory representations for the determination of place of articulation of plosives," Proc. of ICSLP, pages 1643-1646, 1994.

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

[3] B. Lindblom, "On vowel reduction," Speech Transmission Laboratory-Quarterly Progress and Status Report, Stokholm, vol. 29, 1963.

[4] W. Strange, "Dynamic specification of coarticulated vowels spoken in sentence context," J. Acoust. Soc. Am., vol. 85, pages 2135-2153, 1989.

[5] J. D. Miller, "Auditory-perceptual interpretation of the vowel", J. Acoust. Soc. Am., vol. 85, n°5, pages 2114-2134, 1989.

[6] J. Hillenbrand, and R. T. Gayvert, "Vowel classification based on fundamental frequency and formant frequencies," Journal of Speech and Hearing Research, vol. 36, pages 694-700, 1993.

[7] P. Jospa, A. Soquet, and M. Saerens, "Variational formulation of the acoustico-articulatory link and the inverse mapping by means of a neural network," In "Levels in Speech Communication Relations and Interactions," Amsterdam: Elsevier, pages 103-113 ,1994.

[8] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am., vol. 55, n°6, pages 1304-1312, 1974.

[9] E. Zwicker, and E. Terhardt, "Analytical expresions for critical-band rate and critical bandwith as a function of frequency," J. Acoust. Soc. Am., vol. 68, n° 5, pages 1523-1525, 1980.

[10] G. Fant, "Speech sounds and features," MIT Press, Cambridge MA, 1973.

[11] J. Makhoul, "Linear prediction: a tutorial review," Proc. IEEE, vol. 63, pages 561-580, 1975.

[12] M. Mrayati, R. Carré, and B. Guérin, Distinctive regions and modes: a new theory of speech production," Speech Communication, vol. 7, pages 257-286, 1988.

[13] S. Maeda, "Une modèle articulatoire de la langue avec des composantes linéaires," Actes des 10èmes Journées d'études sur la parole, pages 154-162, 1979.

[14] Q. Lin, and G. Fant, "Vocal-tract area-function parameters from formant frequencies," Eurospeech, pages 673-676, 1989.

[15] "SPSS Reference guide," SPSS Inc., 1990.

[16] J. R. Quinlan, "C4.5: Programs for machine learning," Morgan Kaufmann Publishers, San Mateo, California, 1993.