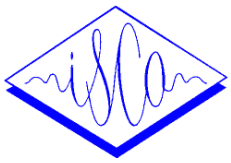

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A COMPARISON OF DIFFERENT ACOUSTIC AND ARTICULATORY REPRESENTATIONS FOR THE DETERMINATION OF PLACE OF ARTICULATION OF PLOSIVES

Alain Soquet [1], *and Marco Saerens*[2]
[1]Institut des Langues Vivantes et de Phonétique
[2]IRIDIA Laboratory
Université Libre de Bruxelles
50 av. F. D. Roosevelt, B-1050 Bruxelles, BELGIUM

ABSTRACT

A problem of long-standing interest in speech processing concerns the most appropriate representation for recognition purposes. The objective of this paper is to compare 6 acoustic and 4 articulatory representations in a task of determination of place of articulation of inter-vocalic plosives. Place of articulation recognition results were obtained based on linear discriminant analysis with the "jackknife" method in which the tokens from each individual are successively removed from the training set, and used as a test set. Systematic comparisons were performed under 3 different sets of conditions depending on whether or not the information about the end of the transition, the transition and the stable part of the vowel are integrated. The LPC cepstrum and two articulatory representations (DRM and Maeda's model) achieved the best recognition rate (86%). However, the 2 articulatory representations appeared to be more stable in terms of inter-speaker variability. The performances of the 7 others representations were found to be significantly lower (74% for formants, and 60% for LPC area).

I. INTRODUCTION

For years, it has been argued that the use of articulatory representation could be advantageous for speech recognition [12][13]. For instance, place of articulation should be more easily identified from articulatory than acoustic representations. The place of articulation of plosives is an obvious example: while the formant transitions surrounding a plosive are known to carry information about the place of articulation [4], the coarticulation effects introduce strong correlations between vowels and transitions. On the other hand, articulatory representations provide direct information about the place of articulation and are therefore highly decorrelated from the surrounding vowels. While the burst is known to provide important information about the place of articulation [10], the present study only focusses on vocalic transitions.

Unfortunately, the computation of the articulatory configuration from the acoustic parameters (the acoustic-to-articulatory inversion) is not a trivial problem. In previous work [3], we developed a tool that realizes this inversion in the framework of an articulatory model, based on the first three formant frequencies. The inversion is performed by a multilayer neural network, trained to approximate the nonlinear mapping from the acoustic space (the first three formant frequencies) to the articulatory space.

II. ACOUSTIC REPRESENTATIONS

The speech signal was passed through a 5 kHz cutoff low-pass filter, and sampled at 10 kHz. The signal was then preamphasized ($1 - 0,95z^{-1}$) 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, a representation very popular in knowledge-based recognition systems.

- **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.
- **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** (LCPST): The LPC cepstral coefficients were derived from the predictive coefficients obtained with an LPC analysis (Atal [1]). As before, we used the first 12 coefficients.
- **Formants** (FORM, BARK, MEL): The formant values were extracted semi-automatically on the basis of the different acoustic representations (in a similar way as described by McGonegal et al. [8]). We used 3 different scales for the frequency axis: Hertz (FORM), Bark [15] (BARK), and Mel [2] (MEL).

III. ARTICULATORY REPRESENTATIONS

Four articulatory representations were selected. The first one is computed from the LPC coefficients. The other three are obtained with acoustic-to-articulatory inversion of three different articulatory models from the formant frequencies (FORM).

- **LPC area** (LAREA): The LPC area functions are computed from the LPC reflection coefficients as suggested by Makhoul [7].
- **DRM** (DRM): The distinctive regions model [9] is an 8 regions acoustic tube with transversal control. The model is derived from acoustical properties of an 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 [6] is an articulatory model derived from X-ray sagital cuts of one speaker. The 7 parameters of the model control the shape of the sagital cuts.
- **Lin Fant** (LF): Lin and Fant's model [5] is a geometrical model with longitudinal control. There are 3 main control parameters (two for the principal constriction, and one for the lips).

The control parameters of the three articulatory models are provided by a neural network realising acoustic-to-articulatory inversion on the basis of the first three formant frequencies (Jospa et al. [3]).

IV. EXPERIMENTS

In order to study the effectiveness of these different representations for the determination of place of articulation of inter-vocalic plosives, 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.

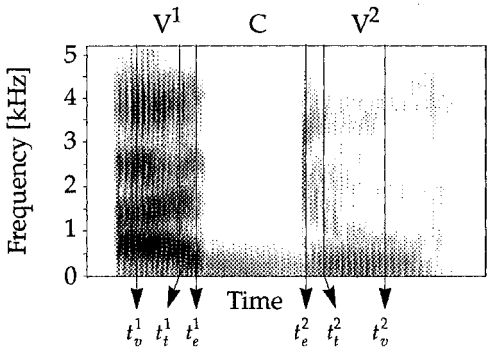

Figure 1: Localisation of the 6 measurement time-instants on one V^1CV^2 token.

The 10 representations have been computed at 3 different time-instants for each transition (see figure 1). The time-instant t_v^1 is located in the stable part of the vowel V^1, t_e^1 at the end of the vocalic transition, and t_t^1 100 ms before t_e^1. Similarly, the time-instant t_v^2 is located in the stable part of the vowel V^2, t_e^2 at the beginning of the transition, and t_t^2 100 ms after t_e^2. The combination of the three measurements per vowel allows us to take into account the transition and/or the context.

V. RESULTS AND DISCUSSION

Place of articulation recognition results were obtained based on linear discriminant analysis with the "jackknife" method in which 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.

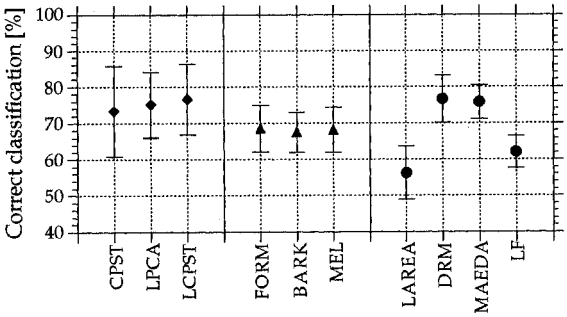

Figure 2: Results of discriminant analysis showing percent of classification of place of articulation and inter-speaker standard deviation with information relative to the vowel V^1 (t_v^1, t_t^1, and t_e^1).

Figure 2 and figure 3 show the results of classification of place of articulation obtained for the ten different representations, respectively using the information relative to V^1 and V^2 (each averaged on 1650 tokens). In general, results obtained for the first vowel are superior to those of

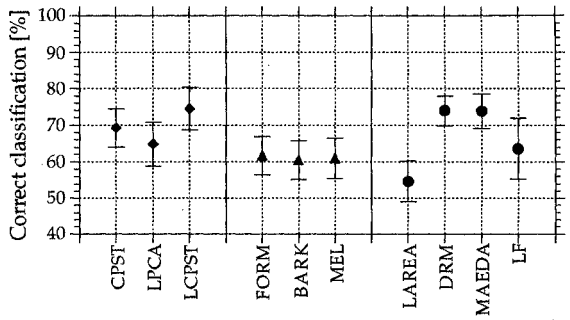

Figure 3: Results of discriminant analysis showing percent of classification of place of articulation and inter-speaker standard deviation with information relative to the vowel V^2 (t_v^2, t_t^2, and t_e^2).

the second one (as in [11]). It can be seen that LCPST gives the best average results followed closely by two articulatory representations (DRM and MAEDA). These two articulatory representations perform about 10 percent better than the three formant-based representations. Since these articulatory representations are computed from the formant values, the nonlinear transformation from the acoustic space to the articulatory space improves the discrimination between the clusters.

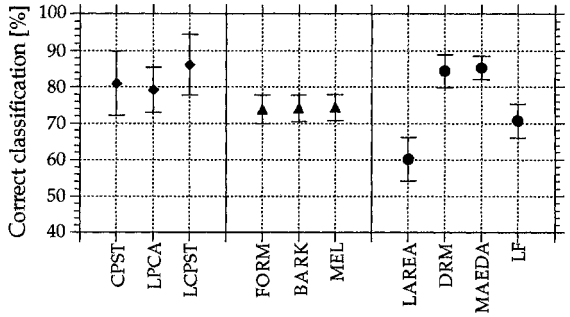

Figure 4: Results of discriminant analysis showing percent of classification of place of articulation and inter-speaker standard deviation with information relative to both V^1 and V^2 (t_v^1, t_t^1, t_e^1, t_v^2, t_t^2, and t_e^2).

The simultaneous use of the information relative to both vowels surrounding the plosive significantly improves the classification results as shown in figure 4 (averaged on 1650 tokens). On average, LCPST and MAEDA achieve the highest scores closely followed by DRM (not significant). However, the variances of the two articulatory based representation are smaller than for LCPST (MAEDA: $p < 0.002$; DRM: $p < 0.1$). The performances of formant-based representations are still 10 percent lower. In all cases, the CPST and LPCA perform less well than the LCPST ($p < 0.01$). The poor performances of LAREA could be due to its limitations and technical problems such as source characteristics, boundary conditions, losses in the vocal tract, and vocal tract length [14]. LF performance is more than 10 percents worse than the two other articulatory-based representations.

Figure 5 shows the influence of the amount of information provided to the discriminant analysis on the classification scores: information at the locus is successively completed with the measurement in the transition and in the stable part of the vowel. The two transitions (either from V^1 or V^2) were treated as being independent (3300 tokens). The trend is similar for the four representations: the performances increase with the amount of information provided to the discriminant analysis.

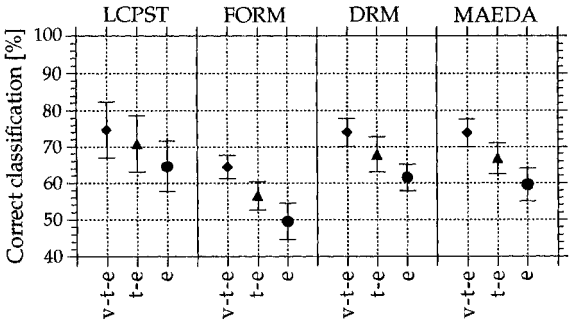

Figure 5: Results of discriminant analysis showing percent of classification of place of articulation for 4 representations depending on which measurements are taken into account for each transition (t_v^i, t_t^i, and t_e^i; only t_t^i, and t_e^i; and t_e^i alone).

In order to study the relationship between the scores and the vowels involved in the transitions, we have successively isolated the transitions of the different vowels [a, œ, i, u, y]. Each set consists of 60 transitions uttered by the 11 speakers (660 transitions). The discriminant analysis was then repeated for each set (with the "jackknife" method on the speakers), for each vowel, and for four representations (LCPST, FORM, DRM, and MAEDA). The results of the mean confusion tables are represented graphically on figure 6.

Depending on the vowel, the classification performances are very different:

- [a] or [œ]: The scores are very high for the 4 representations (more than 90%).
- [i]: The transitions from vowel [i] were found to be hard to classify for the 3 places of articulation. However, the LCPST allows to diminish the misclassification of dentals [t, d].
- [u]: The transitions involving a vowel [u] allow good discrimination between dentals [t, d] versus labials [p, b] or velars [k, g]. The separation between labials and velars was found to be very difficult. Many velars have been misclassified as labial. This trend is somewhat reduced with LCPST.
- [y]: While inferior to the results obtained with vowel [a] or [œ], the performances observed for transitions starting from [y] are satisfactory. The case of velars showed some difficulties (< 75%).

In general, it can be observed in figure 6 that LCPST gives performances markedly superior to those of the 3 other methods. The results of FORM are similar or slightly superior to those of the articulatory representations (DRM or MAEDA).

The comparison of these results with the scores obtained for the 5 vowels simultaneously (from figure 2 to figure 5) raises two observations: **(i)** FORM performs worse than LCPST both on individual vowel transitions and on the whole data set; **(ii)** DRM and MAEDA perform less well than LCPST on individual vowel transitions, but, for the whole corpus, both give results similar to LCPST.

This is surprising since in the first experiments, information about the surrounding vowel was provided (t_v). Moreover, a control experiment allowed us to verify that the vowel is very well identified based on the information computed at t_v. This indicates that the articulatory representations, while less informative than the LCPST, form more compact clusters that are more easily separable with a linear discriminant. Figure 7 shows the scatter plot of the

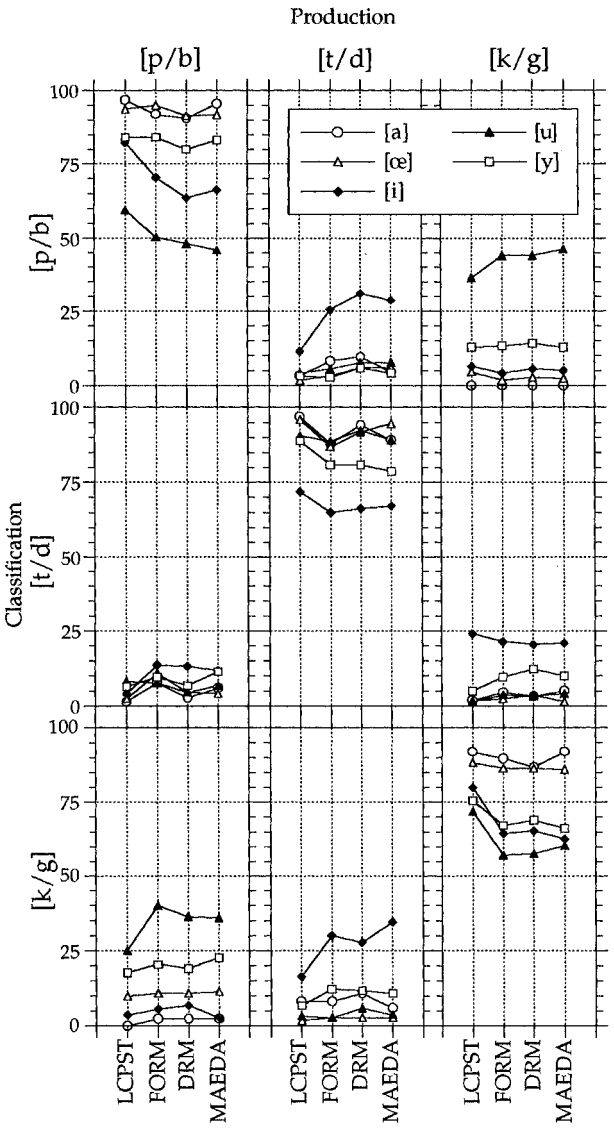

Figure 6: Mean confusion matrix of discriminant analysis for 4 representations depending on the vowels involved in the transistion.

1650 V^1CV^2 on the two first discriminant axes of a linear discriminant analysis with measurements at t_v^1, t_t^1, t_e^1, t_v^2, t_t^2, and t_e^2. We observe a clear separation between the three clusters for the LCPST. The improvement obtained with the transformation from FORM to DRM or MAEDA is also apparent.

Moreover, some results can receive quite satisfactory explanations in the articulatory domain. For example, the [u] has an important constriction at the lips and at the dorsum of the tongue. Therefore, movements towards labial or velar plosives are very small and could be difficult to observe. By contrast, vowels like [a] or [œ] are open in the front cavity. Therefore, movements toward occlusion are wide and easier to detect.

VI. CONCLUSIONS

We compared ten different representations of the speech signal for the classification of the place of articulation of plosives. The LCPST showed the best recognition scores on average. Two articulatory representations (DRM and MAEDA) obtain similar performances on the whole

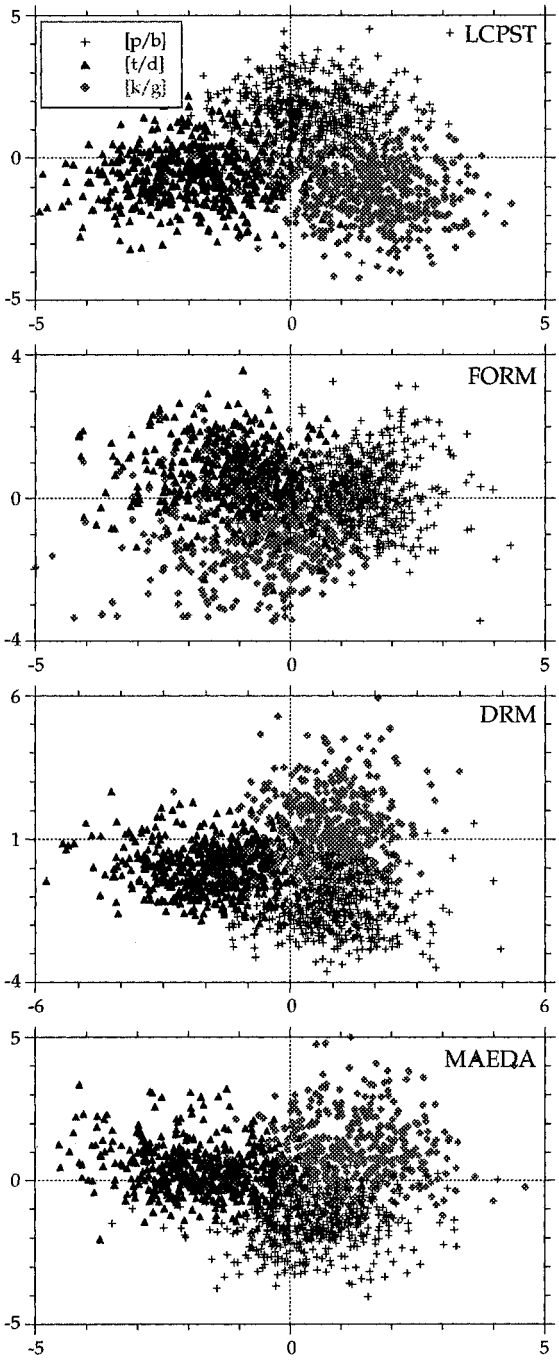

Figure 7: Scatter plot of the 1650 V^1CV^2 tokens on the two first discriminant axes.

corpus, but with lower inter-speaker variability. The performances for the formant representations were found to be significantly less good. Since the articulatory representations are computed from the formant frequencies, the acoustic-to-articulatory transformation enables to increase the linear separability of the clusters. This indicates that the articulatory representation is more suited for place identification that formant values. A next step may be to compute the articulatory representation on the basis of the LCPST, which appeared to perform better than FORM.

ACKNOWLEDGEMENT

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] 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.

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

[3] P. Jospa, A. Soquet, and M. Saerens, "Acoustical sensitivity functions and the control of the vocal tract model," Signal Processing VI: Theories and Applications, J. Vanderwalle, R. Boite, M. Moonen, A. Oosterlinck (editors), Morgan Kaufmann Publishers, pages 319-327, 1992.

[4] D. Kewley-Port, "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," J. Acoust. Soc. Am., vol. 72, n°2, pages 379-389, 1982.

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

[6] 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.

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

[8] C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "A Semi-Automatic Pitch Detector (SAPD)," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 23, pages 570-574, 1975.

[9] 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.

[10] M. E. H. Schouten, and L. C. W. Pols, "Identification of intervocalic plosives consonants: the importance of plosive burst vs. vocalic transitions,"in 10th Congress of Phonetic, Utrecht, pages 464-468, 1983.

[11] D. J. Sharf, and T. Hemeyer, "Identification of place of articulation from vowel formant transitions," J. Acoust. Soc. Am., vol. 51, n°2, pages 652-658, 1972.

[12] K. Shirai, and M. Honda, "Estimation of articulatory motion from speech waves and its application for automatic recognition," Spoken Language Generation and Understanding, J. C. Simon (ed.), D. Reidel Publishing Company, pages 87-99, 1980.

[13] K. Shirai, T. Kobayashi, and J. Yazawa, "Estimation of articulatory parameters by table look-up method and its application for speaker independant phoneme recognition," Proceedings of the Int. Conf. on Acoustic, Speech and Signal Processing, Yokyo, pages 2247-2250, 1986.

[14] H. Wakita, "Estimation of vocal-tract shapes from acousticl analysis of the speech wave: The state of the art," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, n°3, pages 281-285, 1979.

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