RECOGNITION OF FRENCH VOWELS BY EXPERT SYSTEM SERAC

Anne BONNEAU

Centre National d'Etudes
des télécommunications
22300 Lannion (France)

Mario ROSSI

Institut de Phonétique
13100 Aix-en Provence (France)

ABSTRACT

This paper concerns two methods aiming to the automatic recognition of French vowels in continuous speech. The first part presents the results obtained by an algorithm based on the detection of context- and speaker-independent acoustic cues for the fine identification of the vowels. The second part concerns the preliminary results obtained for the detection of the features open/close and front/back, by context-independent cues and partially speaker-dependent cues (the frequency ranges on which certain rules operate are adapted to the sex of the speaker). The limits of the two methods are discussed. It is suggested that recognition of the vowels should be performed using a mixed strategy: an "invariant feature" recognition module, to classify the vowels, followed, for each vowel class, by a specific module which would partially be speaker- and context- dependent.

INTRODUCTION

The detailed recognition of vowels independently of the context and of the speakers is difficult in languages like French which has a rich vocalic vowel system (See Fig.1 the French vocalic triangle). French is generally considered as having four distinctive degrees of opening; nasality is distinctive; it has a series of front unrounded vowels and a series of front rounded vowels.
Speaker- and context-independent cues can however be used for the recognition of the most robust features of the vowels, allowing a gross classification into large vocalic classes. The description of the words by gross features is useful (at least from a computational point of view) to access the lexicon and to select a subset of words to be later verified against the signal. Such a scheme involves a mixed strategy for the recognition of the vowels. After the selection of a subset of vowels sharing the same gross feature(s), done by the use of context- and speaker- independent rules (related to the existence of invariant cues), a vowel in the subset is choosen by a specific module which uses speaker- and context- dependent knowledge. Such a mixed strategy was suggested to us by a careful examination of the results (i.e. vowel confusions) done the algorithm developped by Rossi. The algorithm and its evaluation will be presented in the first part in this paper. At the present state of knowledge, it is not really feasible to identify all the vocalic features by speaker- and context independent rules. The second part deals with the detection of the open/close and front/back features.

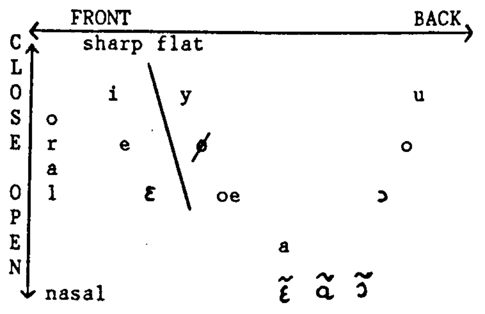

Figure 1: French vocalic triangle

I. VOWEL RECOGNITIONBY ROSSI'S ALGORITM

I.1 Algorithm and results

Speech is first process using a fourteen channel vocoder. The rules apply to the central region of the vowel. Recognition is done on a binary basis, using a tree-type structure. The main acoustic parameter is the vector of energies in each channel of the vocoder. The cues generally make comparisons between the energy levels of two specific frequency ranges in the spectrum.
For example, one of the rules is:

if EK1 >= (EK3 + EK4) then OPEN 1 = TRUE

where EK1, EK3 and EK4 represent the energy level in the first, fourth and fifth channel. The rules have been formulated by the study of 320 CVCV isolated words.
For conveniency, for testing, the rules have been implemented in the SERAC system, an expert system developped at CNET in connection with the Artificial Intelligence group.

The performance of the system has been evaluated on 20 sentences (by two male speakers), 50 connected numbers (by two male speakers) and 300 isolated numbers spoken (by six male speakers). For eacn vowel, one to three candidates are proposed. In other words, the list of candidates for each vowel never contains more than three hypotheses. In average, in 75 percent of the cases, the right vowel is one of the candidates included in the list. In 52 percent of the cases, the first candidate is the right solution (See Table 1).

TEST-CORPUS	SPEAKERS males only	LIST OF CANDIDATES % CORRECT	FIRST CANDIDATE % CORRECT
.SENTENCES (20)	2	79	55
.CONNECTED NUMBERS (50)	2	76	53
.NUMBERS (300)	6	72	50
TRAINING CORPUS: 320 CVCVC WORDS			

Table 1: Percentages of correctly recognized vowels

I.2 Discussions

Confusions occur mainly between:

(1) the back close vowels /u/ and /o/;
(2) /oe/ and its nasal counterpart /ɛ̃/;
(3) /a/ preceded by /k/, and /oe/ preceded by /r/;
(4) the three nasal vowels /ɛ̃/, /ɑ̃/ and /ɔ̃/.

Some of the errors are therefore probably due to the fact that the mid region is not stable and includes transitional movement, and to a lack of information in the very low frequencies. We think also that binary cues are not well adapted at this stage of recognition, and Klatt [4] has already spoken of the undesirability of forcing an early decision.

Although the algorithm passes directly from the cues to the vowels without a clearly defined intermediary level which would be a feature recognition module, it is interesting to evaluate the confusions appearing between vowels of opposite classes: between open and closed vowels, between front and back vowels, or oral and nasal vowels, etc...
Let us do a number of remarks on the assignement of a degree of opening and of a degree of backing to some of the vowels.
(1) The degree of opening of mid vowels -/e,ɛ,o,ɔ,ø,oe/- is not often easy to determine. In some cases, the contrast between the /e/ and /ɛ/, /o/ and /ɔ/, and /ø/ and /oe/ is distinctive. In many cases, the speaker tends to use the open vowels (/ɛ/,/ɔ/,/oe/) when there are embedded in closed syllables; on the contrary, they tend to close (/e/,/o/,/ø/) them in open syllables. Further more, despite such tendencies, it is often the case that the ear is not able to decide whether or not the vowel is open or close. Therefore the error rates presented in this paper do not take into account errors arising between the mid vowels. Two degrees of opening only (close and open) are considered.
(2) We adopt the same restrictions for the vowels /a/, /ɛ̃/, /oe/,/ø/, which were not a priori classified on the front-back axis.
The results, detailed in Table 2, show that there is a few number of confusions between open and closed vowels and between front and back vowels (with the restrictions explained above).

	F	B		O	C		OR	N		FL	S
F	95	3	O	99	4	OR	82	36	FL	43	89
B	5	97	C	1	96	N	18	64	S	57	11

Table 2: CONFUSION MATRIX BETWEEN VOCALIC CLASSES
Only the first vowel candidate is taken into account. The horizontal axis represents the recognized vowel class. The results are given in percentages.
F: front, B: back, O: open, C: close,
OR: oral, N: nasal, FL: flat, S: sharp

To summarize, the vowel recognition module, as developped by Rossi and presented here shows an ability to identify half of the vowels and to correctly classify all the vowels in the main vocalic classes. Context-sensitive rules and speaker adaptation seem necessary to perform finer vocalic distinctions (since there are 48 percent of errors).

From these results, we suggest a recognition process which would be:
(1) relatively speaker- and context-independent for the robust features recognition. By relatively, we mean that we do not exclude a-priori a separation between large classes of speakers - such as men, women-, or to introduce in some cases contextual rules.
(2) speaker- and context-dependent for the detailed recognition of vowels.

We are currently working on a feature recognition module. We present below our methodology and our preliminary results.

II. FEATURE RECOGNITION

The methodology remains the same as that adopted previously by Rossi with two exceptions which take into account our previous remarks:
(1) the cues are only evaluated on a single spectrum sampled in the middle of the vowel in order to minimise the influence of context or that of a poor delimitation of the boundaries of the vowel.
(2) the cues are no longer binary.

The training corpus is made of 160 logatoms of the CVCVCVC-type, where the three vowels and the three consonants are identical-, and V represents 13 French vowels and C the 16 French consonants. The data have been automatically segmented using a segmentation module. Since automatic segmentation is by no way perfect, some of segmentation errors have been manually corrected. The tests were carried out on two speakers, one male and one female.

We have tested about a dozen of rules for the open/close and front/back distinctions. Statistical tests were used to select the most discriminatory ones. Each candidate is given with a confidence score which falls between 0 and 1, according to the values of the cue.
As said previously, the degree of opening of the mid vowels is not determined a priori. These vowels are not taken into account during tests on opening cue validity and, consequently, during the evaluation of the discrimination rate of this cue. During the recognition phase, the mid vowels are automatically classified by the program as either open or closed depending on whether their values values match those of open or closed vowels: this method gives an objective criterion for distinguishing between these allophones. The same strategy is adopted for /a, ɛ̃,oe,ɸ/ which are not a priori classified on the front-back axis.
We will successively present the results obtained on the training corpus itself and on another corpus.

a) Results on the training corpus.

Figure 2 shows the histograms of the acoustic correlates of the two features: the "open 1" and the "front 1" cues. Each of the two features can be identified by a single cue with an error rate lower than or equal to 3%.

Such results can be further improved in two ways:

- ADAPTING SOME FREQUENCY RANGES TO THE SEX OF THE SPEAKER:
This error rate can be brought down by 1% by adapting the frequency range considered by the rule for the "front 1" cue, depending on whether the speaker is a male or a female.

- ADDING A RULE FOR THE /i/-/u/ DISTINCTION:
For certain vowels which sometimes have a weak second formant (front /i/ and back /u/), or a very low one (/u/), the "front 1" cue isn't always well correlated to their place of articulation. A secondary cue adapted to the identification of /i/ and /u/, allows the elimination of a large number of incertainties or errors between these two vowels. In order not to lower the scores obtained with the first cue, only the values of the second cue permitting a sure identification of the feature -i.e when the confidence score is maximal- are used.

To summarize, three cues are enough to identify the open/close and front/back features with an error rate of 1% on the training corpus: a open/close cue, a front/back cue, adapted to separate two set of values for males and females, and another front/back cue only used in cases of certitude.

b) Preliminary results on the test-corpus.

The corpus is made of numbers, the tests were carried out on seven men and seven women. We suggest two ways to evaluate the performances of our set of rules:
- the error rate made when the most probable candidate is considered.
- the error rate done for the recognition of the vowels for which the confidence score is 1 (i.e maximal). Together with this rate error rate is indicated the number of vowels for which this score is obtained. It is, of course, important to get as many as possible vowels with maximal identification score, and to make the least possible errors on those vowels which have obtained maximal score.

FEATURE	ERROR RATE		NUMBER OF CANDIDATES WITH T=1
	T>.5	T=1	
FRONT/BACK	1	0	80
OPEN/CLOSE	1	0	40

Table 3: RECOGNITION RATES FOR FEATURES
T represents the confidence score.

The results confirm the results obtained on the training corpus (see table 3). The errors on the open/close feature only concern the /u/-vowel in the number "douze" (/duz/): back /u/ is identified as front. The same vowel is also responsible for the rather weak number of candidates for open vowels given with a maximal confidence score. We are presently looking for simple solutions to solve the /u/-problem. The great articulatory variations of /u/ have already been noted in dental context in French as well as in other languages [5]. The satisfactory results obtained for the front/back cue let us hope that it will be as effective on larger corpus and for a greater number of speakers.

CONCLUSION

We have proposed in this paper à mixed strategy for the recognition of French vowels : speaker- an context-independent for the recognition of the open/close and front/back features, and speaker- and context-dependent for the recognition of vowels. The aim of the features recognition module is to perform a reliable, prior classification of the vowels. This module can be used independently, principally for accessing the lexicon, or can be connected to vowel recognition modules. On a corpus made of numbers, spoken by 14 speakers, we obtain an error rate of 1% for the recognition of the open/close and front/back features and no error is made on candidates given with the highest confidence score. We may therefore conclude that our algorithm is very reliable. We are presently testing the module on a larger corpus and on a greater number of speakers.

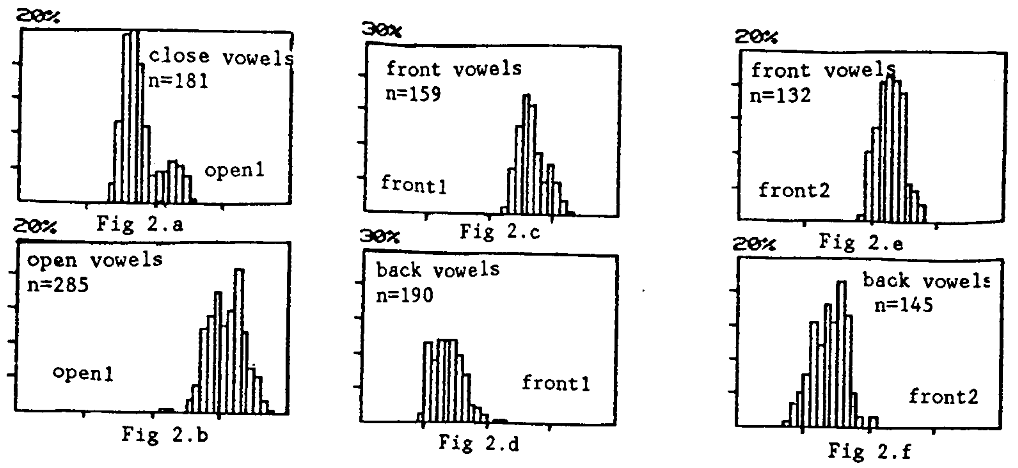

Figure 2: Frequency distribution of acoustic cues values as a function vowel class (results from the training corpus)
n: number of vowels
open1: an acoustic correlate of the open/close feature
front1: an acoustic correlate of the front/back feature
front2: an acoustic correlate of the front/back feature, the same as front1 but with a higher frequency range (adapted for female speaker)

"open1" calculates the difference between EK1 and (EK3 +EK4), EKi represents the level of the energy in the ith channel.

"front1" calculates the difference between (EK6 + EK7) and (EK4+ EK5)

Fig 2.a, Fig 2.b: two speakers (1 M + 1 F)
Fig 2.c, Fig 2.d: male speaker
Fig 2.e, Fig 2.f: female speaker

REFERENCES

[1] D.W. Shipman, V.W. Zue, "Properties of large lexicons. Implications for advanced isolated word recognition systems", Proc. IEEE ICASSP, Paris, 1982.

[2] G. Adda, M. Eskenazi, P.E Stern, "Reconnaissance de grands vocabulaires: utilisation et evaluation de traits grossiers" Journees d'Etudes sur la Parole, Aix-en-Provence, 1986.

[3] A. Bonneau, M. Mercier, M. Gerard, M. Rossi, "Decodage acoustico-phonetique a l' aide du systeme expert Serac-Iroise" Journees d' Etudes sur la parole, Aix-en-Provence, 1986.

[4] D.H. Klatt, "Models of phonetic recognition I: issues that arise in attempting to specify a feature-based strategy for speech recognition", Proc. Montreal Symposium on speech recognition, 1986.

[5] K. Shirai, T. Kobayashi, J. Yazawa, "Estimation of articulatory parameters by table-look method and its application for speaker independent phoneme recognition", Proc. ICASSP, 1984.