RECOGNITION OF FRENCH VOWELS BY EXPERT SYSTEM SERAC

Anne RON THE
Centre National d’Etudes des telecommunications
22500 Lannion (France)

Bario ROSSI
Institut de Phonetique
13100 Aix-en-Provence (France)

ABSTRACT

This paper concerns two methods aiming to the automatic recognition of French vowels in continuous speech. The first presents the results obtained by an algorithm based on the detection of context- and speaker-independent acoustic cues for the identification of the vowels. The second part concerns the preliminary results obtained for the detection of the features open/closed and front/back, by context-independent cues and partially speaker-dependent cues (the frequency range on which certain rules operate is 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 description of vowels independently of the context of the speaker 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 a very distinctive degree of opening; nasality is one of the four distinctive degrees of opening; nasality is the only distinctive degree of opening; it has a series of consonants and vowels and a series of consonants and nasals.

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 for a computational point of view) to access the lexicon and to select a subset of words to be verified against the signal. Such a feature involves a mixed strategy for the recognition of the vowels. After the selection of the set of vowels sharing the same gross feature(s), the use of context- and speaker-independent rules (related to the existence of invariant consonant-vowel in the word) is necessary to classify the vowels. Each of these modules was suggested to us by a careful examination of the results (i.e. speaker and context independence). The second part deals with the detection of the open/closed and front/back features.

1. VOWEL RECOGNITION BY ROSSI’S ALGORITHM

1.1. Algorithm and results

Speech is first processed using a fourteen channel vocoder. The rules apply to the central region of the vocalic system in the range between two specific frequency ranges in the spectrum. By example, one of the rules is:

if [E] or [ε] then [ε] or [ɛ]

where [E] and [ɛ] are the energy level of the first, second, and fifth channel. The rules have been formulated by the study of French and English words.

For convenience, for testing, the rules have been implemented in the SERAC system, an expert system developed at CNET in connection with the Artificial Intelligence group.

The performance of the system has been evaluated on 20 sentences (by two male speakers) and 300 isolated numbers spoken (by six male speakers). For each vowel, two or three candidates are proposed. In other words, the list of candidates for each vowel never contains more than three hypotheses. In average, 75 percent of the cases, the right vowel is one of the candidates included in the list. In 25 percent of the cases, the first candidate is the right solution (see Table 1).

Table 1: Percentages of correctly recognized vowels

<table>
<thead>
<tr>
<th>SENTENCES</th>
<th>70</th>
<th>90</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONNEXION NUMBERS</td>
<td>6</td>
<td>7</td>
<td>50</td>
</tr>
</tbody>
</table>

1.2. Discussions

Confusions occur mainly between:

(1) the back close vowels /u/ and /o/;
(2) /o/ and /a/ nasal counterpart /o/;
(3) /u/ preceded by /a/ and /o/ preceded by /e/;
(4) the three nasal vowels /o/ and /a/ and /e/.

Some of the errors are therefore probably due to the fact that the mid region is not stable and includes transitional movements, and to a lack of information in the very low frequencies. We think also that highly nasaled vowels are not well adapted at this stage of recognition, and Blatt has already spoken of the understandability of forcing an early decision.

Although the algorithm passes directly from the vowels 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.

For example, one of the rules is:

if [E] or [ε] then [ε] or [ɛ]

where [E] and [ε] are the energy level of the first, second, and fifth channel. The rules have been formulated by the study of French and English words.

For convenience, for testing, the rules have been implemented in the SERAC system, an expert system developed at CNET in connection with the Artificial Intelligence group.

The performance of the system has been evaluated on 20 sentences (by two male speakers) and 300 isolated numbers spoken (by six male speakers). For each vowel, two or three candidates are proposed. In other words, the list of candidates for each vowel never contains more than three hypotheses. In average, 75 percent of the cases, the right vowel is one of the candidates included in the list. In 25 percent of the cases, the first candidate is the right solution (see Table 1).

Table 2: CONFUSION MATRIX BETWEEN VOCALIC CLASSES

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: CONFUSION MATRIX BETWEEN VOCALIC CLASSES

Only the first vowel candidate is taken into account. The horizontal axis are the recognized vocalic class. The results are given in percentage.

F: front, B: back, O: open, C: close, OH: oral, NO: nasal, FL: flat, S: sharp

degrees of opening only (close and open) are considered.
(2) we adopt the same restrictions for the vowels /u/, /o/, /e/, /a/, /i/, /ii/, /y/, which were not a priori classified on the front-back axis.

The results, detailed in Table 2, show that there is a number of confusion between open and closed classes, between front and back vowels (with the restrictions explained above).

To summarize, the vowel recognition module, as developed 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 are necessary to perform four vocalic distinctions (since there are four percent of errors).

From these results, we suggest a recognition process which would be:

(1) a speaker- and context-dependent for the robot features recognition by relatively, we mean that the robot should take into account many parameters.

(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 and the CVOCDE type, to what the French vowels and the three consonants are adapted to the French.

(2) the cues are no longer binary.

The training corpus is made of 160logarnoms of the French words and English words and the three
c consonants are adapted to the French.

The data have been expressed in a CNET module. Since acoustic segmentation is no way necessary to a correct classification, the tests were carried out on two speakers, one male and one female.

Se 53.5.1 282

Se 53.5.2 283
We have tested a dozen rules for the open/close and front/back distinctions. Statistical tests were used to select the nine discriminative ones. Each candidate is given with a confidence score which falls between 0 and 1, according to the values of the criterion.

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 close vowels 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 match those of open or closed vowels; this method gives an objective criterion for distinguishing these allophones. The same strategy is adapted for /i, e, o/ which are not a priori classified on the front-back axis.

We will successively present the results obtained for the open close and front-back axes. The results concern the training corpus itself and on another corpus.

a) Results on the training corpus.

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

Such results can be further improved in two ways:

- ADAPTING SOME FREQUENCY RANGES TO THE SEX OF THE SPEAKER.
- ADAPTING A RULE FOR THE /u/-/u/ DISTINCTION.

For certain vowels, which sometimes have a weak second formant (front /a/ and back /u/, or a very low one /u/), the "front" cue isn’t always well correlated to sharp pitch fluctuations. A second /u/ cue adopted to the identification of /a/ and /i/ allows the elimination of a large number of incorrects or errors between these two vowels. In order not to lower the second cues used with the first cue, only the values of the second cue permit fitting a more identification of the feature /i/ when the confidence score is minimal.

To summarize, three cues are enough to identify the open/close and front/back features with an error rate of 3% on the training corpus: a open/close cue, a front/back cue, and a second front/back cue used only in cases of criteria.

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 which the confidence score is 1 (i.e. maximal)." Together with this rate error rate which is indicated the score of vowels for which this score is obtained. It is, of course, important to get as many as possible on the maximal identification score, and to make the least possible errors on those vowels which have obtained maximal score.

### Table 3: Recognition Rates for Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Error Rate</th>
<th>Number of Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRONT/BACK</td>
<td>1.0</td>
<td>80</td>
</tr>
<tr>
<td>OPEN/CLOSE</td>
<td>1.0</td>
<td>60</td>
</tr>
</tbody>
</table>

The results confirm the results obtained on the training corpus (see Table 3). The errors on the open/close feature only concern the /a/ /u/ vowel in the number "douze" / dial/; / back /u/ is identified as / front. The same vowel is also responsible for the other vowel number of candidates for open vowels given with a maximal confidence score are presently looking for simple solutions to assess the /u/-/u/ problem. The generalization of /u/ have already been noted in dental context in French as well as in other vowel features. The satisfactory results obtained for the front/back cue let us hope that it will be an effective on larger corpus and for a greater number of speakers.

### Conclusion

We have proposed in this paper a mixed strategy for the recognition of French vowels: speaker-independent for the recognition of 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 10 speakers. We obtain an error rate of 3% for the recognition of the open/close and front/back features and an error rate on candidates with a maximal confidence score. We may therefore conclude that our algorithms is very reliable. We are presently testing the module on a larger corpus and on a greater number of speakers.

### References