DURATION OF CONSONANT CLUSTERS IN FRENCH: AUTOMATIC DETECTION RULES

Yukihiro Nishinuma, Danielle Duez, and Chantal Paboudjian
Institut de Phonétique, Université de Provence
CNRS U.A.-261, 13621 Aix-en-Provence, France

Abstract
This study describes how to automatically detect clusters of 2 consonants in a spoken French word recognition system. A set of rules was deduced from the statistic analyses carried out on the 4 corpora recorded by 10 subjects. Five relevant parameters were extracted, namely: 1) voice feature and 2) mode of articulation on the first half of the consonant segment, 3) duration ratio between vowel and consonant segment, 4) duration of consonant segment, and 5) position (prevocalic, intervocalic, postvocalic). This phonetic approach was examined on the GRECO-BDSONS, public test corpus. Our classification was 90% correct using the standard values of the duration parameters provided by our data base.

INTRODUCTION
Research dealing with the duration of consonant clusters is not very common, and mainly concerns Swedish, English, Arabic, and French (1 - 11). These studies have been aimed at assessing the compression effect undergone by the vowel and/or the consonant when one or two consonants are added, and in some cases, at defining rules to be applied to speech synthesis.

The purpose of this research is to define rules for automatically detecting clusters of two consonants in continuous spoken French. The study deals with the duration and constituents of consonant clusters, and was conducted in an essentially pragmatic perspective.

EXPERIMENTAL PROCEDURE
Four corpora were used. The first consisted of 50 two-syllable words occurring in the same carrier sentence. The target consonant clusters were placed in the intervocalic context, which was limited to the vowels /a/ and /i/. The second corpus was designed solely for examining the duration of the fricative /s/ followed by one of the stops /p,t,k/, sonorants /l,R/, or nasals /m,n/. The third corpus was complementary to the first and contained 250 words. The voiced or unvoiced stop was examined, although the vocalic context was extended to include /a/. The target segments were placed at the beginning of the first or second syllable in the word. The fourth corpus was composed of one-syllable or two-syllable words with the target segments in word-final position.

Ten speakers participated in the experiment. The recoding was done in an anechoic room. The speakers were asked to read five times the corpora in a "neutral" fashion and at a normal speech rate.

Segmentation was done in compliance with the rules defined by Autesserre and Rossi (12) on a signal editor (PDP 11/24).

RESULTS
Most of the recordings were analyzed acoustically and statistically. The measures pertaining to the syllable were coded and saved in a small data base designed for this project. In order to define a variable that would distinguish isolated consonants from consonant clusters, the relationships between consonant durations (C1 vs C2), vowel duration, and syllable duration were examined. The ratio of consonant segment duration to vowel duration can be used to express the relative length of the consonant segment in
the syllable. If the consonant segment is shorter
than the vowel, this ratio will be less than 1, and
vice versa. Approximately speaking, this ratio
was much less than 1 for isolated consonants
(ranging from .6 to .8 in the initial and medial
positions), but was very often above 1 for
consonant clusters (see Table 1, column CC/V).
Note that there is a slight positive correlation
between consonant cluster duration and this ratio (r = .564).

DETECTION PROCEDURE

Definition of rules

Spectral information (cues or acoustic properties) is assumed to be known, and is used
for segmenting (into vowel and consonant segments), detecting voicing, and specifying the
features of the articulation mode.

To be considered satisfactory, the processing should separate consonant clusters
from isolated consonants. The number of necessary computations should be reduced to the
number of lexical accesses.

On the basis of the statistically significant
differences, seven macro-classes of consonant
clusters were predefined. When applied, the
rules should be able to detect these seven macro-
clusters. Place of articulation is not taken into
account in these rules because of its overlapping
standard deviation between consonants of a
given class. The expected outputs of the rules
are: 1) p t k + , 2) b d g + , 3) f s j + , 4) v z J + ,
5) l r + p t k , 6) l r + b d g , 7) l r + f s j .

A rule is made up of five logical tests
concerning:
- The position of the consonant cluster in the
  word.
- The voice feature of C1.
- The mode of articulation of C1.
- The total duration of the consonant segment.
- The ratio of the consonant segment to the
  adjacent vowel.

The positions defined are: initial, intervocalic, and final. C1 voicing information (%
voiced) is assumed to be known. By analyzing the
first half of the consonant segment, the mode of
articulation can be specified: stop, fricative,
sonorant, nasal.

The last two tests in the rule concern the
detection thresholds defined on the basis of the
statistical results obtained with our data base
data (see Table 1). The absolute reference
threshold, called the critical consonant cluster
duration, was computed from the mean duration
M and the standard deviation S of the macro-
class in question, i. The threshold T is defined in
the following manner:

\[ T(i) = M(i) - 1 \times S(i) \]

This threshold is equal to the minimal
reference duration (expressed in ms) of a given
macro-class.

In order to neutralize across-speaker
variability, a relative threshold computed in the
same way as the absolute threshold was
introduced, i.e. the ratio of the duration of the
consonant segment to the duration of the vowel,
called the critical consonant cluster ratio. For a
given intervocalic consonant segment \((VC1C2V)\),
the vowel to the right of the segment is used.
This threshold can be used to distinguish a

<table>
<thead>
<tr>
<th>Context</th>
<th>C1</th>
<th>C2</th>
<th>CC/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>C1</td>
<td>C2</td>
<td>m</td>
</tr>
<tr>
<td>CC/#_V</td>
<td>ptk</td>
<td>lr</td>
<td>45</td>
</tr>
<tr>
<td>CC/#_V</td>
<td>bdg</td>
<td>lr</td>
<td>85</td>
</tr>
<tr>
<td>CC/#_V</td>
<td>fs</td>
<td>ptklrm133</td>
<td>31</td>
</tr>
<tr>
<td>CC/#_V</td>
<td>vz</td>
<td>85</td>
<td>27</td>
</tr>
<tr>
<td>CC/V_v</td>
<td>84</td>
<td>26</td>
<td>62</td>
</tr>
<tr>
<td>CC/V_v</td>
<td>71</td>
<td>14</td>
<td>142</td>
</tr>
<tr>
<td>CC/V_v</td>
<td>85</td>
<td>33</td>
<td>96</td>
</tr>
<tr>
<td>CC/V_v</td>
<td>ptk</td>
<td>lr</td>
<td>168</td>
</tr>
<tr>
<td>CC/V_v</td>
<td>bdg</td>
<td>lr</td>
<td>205</td>
</tr>
<tr>
<td>CC/V_v</td>
<td>fs</td>
<td>lr</td>
<td>170</td>
</tr>
<tr>
<td>CC/V_v</td>
<td>vz</td>
<td>161</td>
<td>23</td>
</tr>
<tr>
<td>CC/V_v</td>
<td>ptk</td>
<td>198</td>
<td>26</td>
</tr>
<tr>
<td>CC/V_v</td>
<td>bdg</td>
<td>186</td>
<td>27</td>
</tr>
<tr>
<td>CC/V_v</td>
<td>fs</td>
<td>188</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 1. Summary of consonant cluster duration
simple stressed syllable (CV or VC) from an unstressed syllable containing a consonant cluster (C₁C₂V) whenever C (in CV) and C₁ are the same (e.g. "sa#" vs. "sta"). Even if these two syllables have a similar duration, the duration ratio between the vowel and the consonant segment will be very different.

```c
if
    position_in_word = initial;
    CI_voicing = voiced;
    CI_mode_of_articulation = stop;
    critical_ratio_Cseg_V >= .64;
    critical_Cseg_duration >= 70;
then
    number_Cs_in seg = 2
    ConCluster_context = "ptk+something";
end
```

**Figure 1.** Example of a detection rule

In all, seventeen rules were defined. The rules are applied to each suspected syllable nucleus, from left to right. **Figure 1** illustrates one of the rules.

**Application of the rules**

The rules were applied to 144 words from the GRECO corpus (BDSONS #ACCO1-4) read by five speakers. The GRECO corpus includes 204 one-syllable or two-syllable words containing consonant clusters. Those containing "semi-consonants" were excluded here because no experimental data is available on them. One hundred and fifty VCV sequences were added to this corpus (5 subjects x 30) so as to re-establish the numerical balance between VCV and VCV, since the GRECO corpus is void of VCV sequences. The items tested contained a total of 1261 consonant segments, 430 simple consonants and 831 consonant clusters.

The corpus was segmented on the basis of the criteria defined in the experimental procedure. The cumulative score obtained by applying the rules to the data obtained for all five subjects was 749/831 (= 90.13%) of the sequences were thus correctly classified into the six consonant cluster classes (classes 5 and 7 were grouped together due to the insufficient number of observations).

**Table 3** gives the confusion matrix that summarizes the performance of the phonetic rules for the data as a whole (all 5 subjects taken together). Consonant clusters that begin with a stop are the most frequent in French (Auberge et al., 1988); indeed, they make up half of the consonant clusters in the corpus and are detected the best. Consonant clusters with /p,t,k; b,d,g/ as C₁ were detected 411 times (94.70%) as such, and only 24 times as isolated consonants. This score is quite a bit higher than that obtained for fricatives (84.05%) and consonant clusters beginning with a sonorant (85.89%).

**Table 3.** Overall results using the phonetic rules (VZ = voiced, IVZ = unvoiced)

<table>
<thead>
<tr>
<th>OUTPUT</th>
<th>pk+</th>
<th>fof+</th>
<th>bdg+</th>
<th>vVZ</th>
<th>lv+VZ</th>
<th>lvVZ</th>
<th>C</th>
<th>sum</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>pk+</td>
<td>259</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td>274</td>
<td>94.5%</td>
</tr>
<tr>
<td>fof+</td>
<td></td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22</td>
<td>112</td>
<td>80.4%</td>
</tr>
<tr>
<td>bdg+</td>
<td></td>
<td>152</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>160</td>
<td>85.0%</td>
</tr>
<tr>
<td>vVZ</td>
<td></td>
<td></td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>51</td>
<td>92.2%</td>
</tr>
<tr>
<td>lv+VZ</td>
<td></td>
<td></td>
<td></td>
<td>55</td>
<td></td>
<td></td>
<td>14</td>
<td>69</td>
<td>79.7%</td>
</tr>
<tr>
<td>lvVZ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>146</td>
<td>19</td>
<td>186</td>
<td>88.5%</td>
<td></td>
</tr>
</tbody>
</table>

| output | 90.1% |

Are the results obtained with this phonetic model satisfactory? How efficient is this model? In an attempt to answer these questions, a discriminant analysis was performed on the same data. The numeric variables used were the values of the detection rule terms; the category variables (position, mode of articulation, voicing) were converted into binary numerical variables. Using this multi-variable statistical model, consonant cluster detection scored 96.39%.

The score obtained by the statistical model was 6.26% higher. This can be explained in part by the fact that the computed variable distances were completely optimized in the discriminant analysis, whereas the phonetic model thresholds, which were extracted from the data base, were not optimized. When the thresholds were corrected to take the subject's speech rate into account, the phonetic model scored much higher. For example, the score for subject 2 was high as 95.27% (161/169).
To what extent are the thresholds and features defined in this model operational? In order to determine the actual role played by the temporal factor, the phonetic model system was run using two thresholds (ratio of consonant segment to vowel and consonant cluster duration) and the results obtained were compared to the discriminant analysis with two corresponding numeric variables. Without specification of the consonant features, the discriminant analysis model did not classify the consonant clusters (50.61%) as well as the phonetic model (65.10%). These mediocre scores could have been predicted in the light of the bimodal distribution of the data in which single consonants and consonant clusters were taken together. More than one third of the detection cases are thus based on the voice feature and the $C_1$ mode of articulation feature. The correlation coefficients obtained suggest that the mode of articulation contributes more to detection.

CONCLUSION

In the present study, the validity of incorporating the temporal variable into consonant cluster detection was tested. A data base containing some representative contexts was generated so as to reveal the significant statistical facts describing the isolated consonant and consonant cluster populations.

In our rules, once the position of the consonant segment in the word, the voice feature and the mode of articulation feature of $C_1$ are specified, the absolute and relative thresholds defined for each consonant cluster can be used to correctly classify the consonant clusters into macro-classes. The temporal parameter may thus be used to define automatic consonant cluster detection rules, and is an effective complement to acoustic and distributional

REFERENCES

(2) Benkirane, T., (1982), *Étude phonétique et fonctionnelle de la syllabe en arabe marocain*, Université de Provence (Doctoral dissertation)