STRONG CUES FOR IDENTIFYING WELL-REALIZED PHONETIC FEATURES

A. Bonneau, S. Coste-Marquis and Y. Laprie
CRIN-CNRS & INRIA-Lorraine, Nancy, France

ABSTRACT

We show that there exist cues, called strong cues, which allow some phonetic features to be identified or eliminated with certainty. The use of strong cues brings numerous advantages in ASR, including reliability, and reduction of the search space. With a view to validating our approach, we have defined a first set of strong cues characterizing the place of articulation of French stops. The firing rates of these cues are relatively high. The definition of strong cues can be extended to other features and is useful for different ASR approaches.

1 INTRODUCTION

Recognition techniques generally rely on the use of continuous criteria (as, for example probability density functions for statistical methods). A major disadvantage of this kind of decision is that exemplary realizations of some cues cannot be taken into account and thus cannot lead to definite decisions. Hence, we propose to define strong cues which allow some phonetic features to be identified or eliminated with certainty. Strong cues are both discriminating and well pronounced (according to its realization, a cue can be strong or weak). They must not be confused with main or robust cues.

2 ACOUSTIC CUES

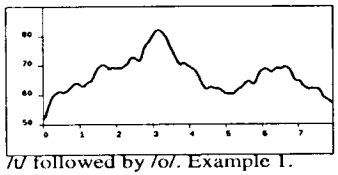
We choose to use context-dependent acoustic cues. For this purpose, we distinguish three classes of vowels: high front vowels (called from now on front vowels), open front and central vowels (called central vowels) and back vowels.

2.1 Description

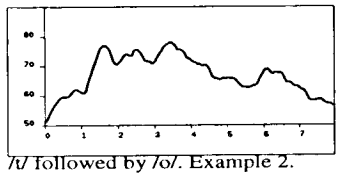
Cues provided by the burst

We use a context-dependent compactness cue to distinguish the velars which are compact from the labials and the dentals which are diffuse. A burst is considered compact if the energy is concentrated around a particular peak. This peak must be the most prominent peak of the spectrum and must be located in a restricted frequency region corresponding to the (context-dependent) region of the spectral maximum of velars. The compactness cue may be more or less pronounced according to the pronunciation of the consonant. We mean that the level of energy concentration may be more or less high and the frequency of the peak more or less close to the expected value. When the cue is very pronounced, the velar place can be identified with certainty.

We use a context-dependent acuteness cue to distinguish dentals from labials. High frequency peaks (above 2200 or 2500 Hz) generally dominate the dental spectrum while low frequency peaks often dominate the labial spectrum. This spectral configuration may not be so clear (or even may not be observed) for all the pronunciations of a consonant. For example, with regard to dental spectra, we know the existence of a peak near the F2 dental locus [1]. This peak sometimes dominates the spectrum. An other example, labials preceding front vowels are often more acute than what is reported in the literature and we frequently observed labial spectral maxima at 2500 or 3000 Hz, which may create confusions with dental consonants. Nevertheless, we believe that, when the dental cue is particularly pronounced, the identity of the consonant is not to be questioned. Figure 1 shows the spectra of two dental consonants, each followed by the back vowel /o/. The two consonants were uttered by the same speaker, in the same sentence, and were both in stressed position. Our stop recognition module, described in [2], detected a strong dental cue for the first example -i.e. in this vocalic context, the identification of /t/ was certain-, while a weak dental cue was detected for the second example -/t/ was the preferred but not the only solution-.



/t/ followed by /o/. Example 1.



/t/ followed by /o/. Example 2.

Cues provided by the transitions

Below are only described CV syllable transitions. The opposite transition trajectories are used for the corresponding VC syllables.

/p,b/. Labialisation lowers F2 and F3 of unrounded vowels. Thus transitions between labials and subsequent unrounded vowels are rising. For rounded vowels, transitions are relatively flat. Since the labial articulation does not require the active participation of the tongue, vowel-to-vowel movement involving this articulator may happen during the articulation of the consonant. As a consequence, if one vowel of a VCV sequence is far less anterior than the other, the expected transition trajectory may not be observed for this vowel.

/t,d/. F2 transitions between vowels and dentals come from a well determined frequency area: the F2 dental locus (around 1500-2000 Hz). Consequently, CV transitions are falling before back vowels (in this context, they constitute a very clear and systematic cue), are falling or flat before central vowels and are flat or rising before anterior vowels.

/k,g/. Cues vary according to the vocalic context. When the vowel is a back vowel, transitions are relatively flat. When the following vowel is a central or a front vowel, F2 and F3 move away (velar pinch). In cases of great coarticulation, F2 and F3 of central vowels, normally spaced, are close together even at the vowel center. Consequently, the velar pinch is not observed.

2.2 Strong cues

As previously observed, a cue can be more or less pronounced according to the realization of a given feature. When a discriminating cue is well pronounced, the feature identification is not only certain but also direct since the detection of other cues becomes useless. We call such a cue a strong cue. Moreover, some spectral configurations are never observed for a given feature. Such a knowledge is valuable for ASR since it allows to eliminate a feature with certainty. We thus propose to define strong preference cues as well as strong exclusion cues. These cues will be used as "anchor points" in the acoustic-phonetic decoding step (see section 4).

Strong cues provided by the burst

Because of the flat spectrum and of the great spectral variability of labials, we do not believe we can define effective strong cues for these consonants.

Context /u, o, ɔ/

Strong preference cue for velars: the energy is concentrated around a prominent peak situated in front of the F2 of the subsequent vowel.

Strong exclusion cue for velars: the lack of a peak in the vicinity of the F2 of the following vowel.

Strong preference cue for dentals: a relatively prominent peak in high frequencies (between 2200 Hz and 4000 Hz).

Context /a, ɛ, œ/

Strong preference cue for velars: the energy is concentrated around a promi-

nent peak situated in the region delimited by the F2 and the F3 of the following vowel.

Strong preference cue for dentals: a relatively diffuse prominent peak in high frequencies (between 2400 and 5000 Hz).

Strong cues provided by the transitions

Context specification allows us to define strong preference cues. For example, a rising CV transition is a discriminating cue for labials followed by a central vowel. Nevertheless, we do not want to define strong preference cues for transitions. The main reason which motivated our choice is that formant tracking is a particularly difficult task at a boundary between vowel and consonant. Nevertheless, it is easier to establish that a trajectory do not correspond to the expected one. If the observed trajectory is drastically opposite to it, the definition of strong exclusion cues is generally possible. Note that strong cues are only considered if the quality of the corresponding acoustic detectors (formant tracking algorithm) is good.

Strong preference cue for velars followed by a central vowel: F2 and F3 come close together.

Strong exclusion cue for labials: F2 of one vowel (in a V-stop-V sequence) takes the opposite direction to that expected, although the other vowel is not more anterior than it.

Strong exclusion cue for dentals: F2 certainly does not come from the dental locus.

3 EXPERIMENTAL RESULTS

3.1 Methodology

We tested the strong cues described on three French corpora. The first corpus (extracted from BDSONS) we used contains isolated words spoken by 5 male speakers. The two other corpora are made of continuous speech. VERLOC was recorded in an office and is constituted of 17 sentences spoken by 16 male speakers (3, 4 or 5 repetitions). The last one contains 22 read sentences made up of stops and vowels, each sentence is pronounced 3 times by 4 male speakers.

The tested items /? stop V/ were extracted from the corpora and hand-labelled. The burst analysis and formant tracking algorithm come from Snorri [3].

3.2 Results

We give the firing rate of strong cues on Table 1 (preference and exclusion cues). We used 758 unvoiced stops and 1769 voiced stops for central vowel context and 610 unvoiced and 933 voiced stops for back vowel context.

	Excl. /p,b/	Excl. /t,d/	Excl. /k,g/	Pref. /t,d/	Pref. /k,g/
p	–	12.5	35.5	–	–
t	1.5	–	41	46.5	–
k	≈ 0.5	18.5	–	–	48.5
b	–	17	44	–	–
d	6.5	–	36.5	23.5	–
g	0.5	27.5	–	–	17.5

Table 1: Firing rates of strong cues for back vowel context (%).

The fact that strong cues alone allow a direct conclusion in more than 40% of cases in the back vowel context validate our approach, even if the back vowel context is undoubtely the most favourable context. Partial results obtained for the central vowel context show that the exclusion cues are only slightly less discriminating than for back vowels (5% lower on average) and thus remain very interesting. However it appears that the burst decomposition algorithm designed for back vowels should be adapted to other vowel contexts in order firing rates of strong cues defined on the burst become higher.

4 ADVANTAGES OF USING STRONG CUES

With regard to the automatic speech recognition, the use of strong cues brings numerous advantages such as the reliability of the information provided, the reduction of the search space, and the possibility to maintain the coherence during the decoding process. Let us develop the two last points.

4.1 Reduction of the search space

If at least one strong cue is detected, the number of acoustic cues and the number of phonetic and lexical solutions are reduced. Indeed, the search for weak cues becomes useless when a preference cue is detected, and limited when an exclusion cue is detected. Taking the decision is then simpler since there is no need to use score combinations, always difficult to turn out. The number of phonetic solutions is reduced: only one solution is proposed when a preference cue is observed, one or several solutions are definitely dismissed when an exclusion cue is detected. Furthermore, using strong cues as confidence islands decreases the search space of the lexical module [4]. For example, if, at the beginning of a word, the dental feature is identified (or dismissed) by a strong cue, the word proposed as the solution must (or must not) begin with a dental consonant.

4.2 Consistency of the decoding process

Strong cues can be used to maintain the consistency of the reasoning during the decoding stage. This strategy has been adopted by the system Daphné [4]. The principle is that strong cues must not contradict one another. Then, in case of conflict between two strong cues, the context in which these cues have been detected has to be questioned. This context includes essentially the segmentation stage, the acoustic detectors (the formant tracker, the burst detector...), and the scope of the coarticulation phenomenon. Let us give a real example we have encountered with the system Daphné. In a VstopV context, the following two strong cues were detected: an exclusion cue for dental stop provided by the transition situated on the left-hand side of the consonant, a strong preference cue for dental stop provided by the burst. This contradiction had to be explained. The system questioned the context and found that the stop closure was relatively long. A new hypothesis, the presence of a stop cluster, removed the contradiction and led to the right solution.

5 CONCLUDING REMARKS

We have shown that, thanks to the use of strong cues, the stop place of articulation can be identified or eliminated with certainty in numerous cases. Strong cues can also be defined for other features, particularly for the place of articulation of fricatives and for the features characterising the manner of articulation. We also believe that the use of strong cues, which brings numerous advantages for knowledge based recognition systems, can be of great interest for systems based on statistical methods.

REFERENCES

1 S. Blumstein and K. Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, 66:1001–1017, 1979.

2 A. Bonneau, S. Coste, L. Djezzar, and Y. Laprie. Two levels acoustic cues for consistent stop identification. In *Proc. of the International Conference on Spoken Language Processing*, volume 1, pages 511–514, Banff (Alberta, Canada), 1992.

3 Y. Laprie and L. Mercier. Un environnement logiciel pour un atelier phonétique. In *Actes des XXèmes Journées d'Étude sur la Parole*, pages 209–214, Trégastel, 1994.

4 S. Coste-Marquis. Interaction between most reliable acoustic cues and lexical analysis. In *Proc. of the International Conference on Spoken Language Processing*, pages 1187–1190, Yokohama (Japan), 1994.