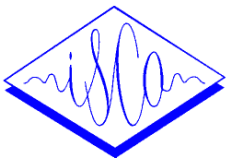

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Two level acoustic cues for consistent stop identification

Anne Bonneau, Sylvie Coste, Linda Djezzar and Yves Laprie
CRIN CNRS & INRIA Lorraine
B.P. 239, 54506 Vandoeuvre-les-Nancy, France

10.21437/ICSLP.1992-164

Abstract

We present an analytical approach for automatic speech recognition based on the detection of acoustic cues. We have defined two levels of cue: "strong" cues both highly visible and efficiently discriminating, and "weak" cues, used in case of absence of "strong" cues. Our strategy based on the two levels of cue has the following advantages:

- it prevents the recognition system from proposing an inconsistent solution
- it permits an early decision when possible thanks to the detection of "strong" cues.

We obtain nearly 50% correct early stop identification with "strong" burst cues; that validates our approach. An operational system is already running for voiceless stops followed by back vowels in continuous speech.

1 Introduction

We present an analytical approach for automatic stop identification for multispeaker continuous speech recognition based on the detection of relevant acoustic cues which aims at maintaining the consistency of reasoning. To identify vowels and consonants, we must deal with at least three problems:

1. the uncertainty of data which results from the unreliability of automatic acoustic detectors (formant tracking algorithm, burst analyser...),
2. the possible absence of some acoustic cues,
3. the fact that the discrimination efficiency of acoustic cues depends on their realization.

In order to take into account those speech peculiarities, we have defined two kinds of cue: "strong" and "weak" cues. If, at the time of their realization, cues are both well pronounced in the sense of (1) and efficiently discriminating, they will be used as "preference" or "exclusion" cues. We call these cues "strong" cues. They aim at allowing stop identification if they do not contradict each other, "preference" cues allow immediate identification, "exclusion" cues eliminate some identification candidates. In the case of absence of "preference cue", the reasoning process relies on less pronounced cues ("weak" cues). Our strategy based on the two levels of cue has the following advantages:

- it prevents the recognition system from proposing an inconsistent solution
- it permits an early decision when it is possible thanks to the detection of "strong" cues.

We present below our stop cues, the implementation of cues provided by burst and finally, a system based on these cues (formant transitions and burst).

2 Stop cues

2.1 Description of cues proposed

Stops are characterized by a total closure in a given place of the vocal tract, the place of articulation. French stops have one of these following three places of articulation; labial (/p,b/), dental (/t,d/), palatovelar (/k,g/). The silence which occurs during the closure is followed by a burst resulting from pressure release. The main stop acoustic events are the formant trajectories at the vowel onset and offset and the burst.

The manner of using cues as well as cues detected on the right-hand side of the stop are detected according to the class of the subsequent vowel in order to take into account the coarticulation phenomenon. Formant transitions on the left-hand side of the consonant are interpreted according to the class of the preceding vowel. We have distinguished three classes of vowel: front vowels, central vowels and back vowels. Classes are defined according to F1 and F2 frequencies.

We adapt well-established knowledge to our recognition strategy: in indicating when the presence or the absence of a cue is certain ("strong preference" (SP) or "strong exclusion" (SE) cue), or when the presence of a feature is not certain ("weak" cue). Note that the quality of detection is taken into account at this stage of the reasoning: an unreliably detected cue cannot become a "strong" cue.

We have chosen reliable cues, that are, from our point of view, cues whose detection is not too difficult, relatively discriminating cues and relatively resistant cues to the different sources of vari-

ation acting in continuous speech. That explains the few uses of F3, used only when necessary because of detection problems or the absence of duration cues which are too sensitive to the sources of variation..., (see also other comments below).

We try to take into account the most frequent coarticulation phenomena in VCV contexts: e.g. for the SE labial cue or the SP palatovelar cue (see below).

We describe now our stop cues in a VCV context. We first list the cues, then we will make some comments. When the context of application of a cue is limited, it is indicated between square brackets.

2.1.1 Acoustic cues provided by the formant transitions

Below, we indicate the cues in CV contexts. The same cues are used in the same VC contexts but in opposite directions.

Weak cue

Labial rising F2.

Dental F2 moves comes from the dental locus (nearly 1,500-2,000 Hz)

Palatovelar F2 and F3 move away [central and front vowels].

Strong preference cue

Palatovelar F2 and F3 at or near the center of the vowel are close together [central vowels].

Strong exclusion cue

Labial F2 of one vowel takes the opposite direction to that expected, although the other vowel is not more anterior to it.

Dental F2 certainly does not comes from the dental locus.

Palatovelar F2 and F3 take the opposite direction of that expected [central and front vowels].

Comments:
Due to formant tracking problems, especially at vowel boundaries, we cannot yet say whether the "strong" cues will be used often.

Labials Labialization lowers formant frequencies but this lowering is not always observed when another factor of coarticulation interacts. If one vowel of the / V labial V / context is far more anterior than the other and if the two vowels are coarticulated (labial consonants require only the lips), F2 of the less anterior vowel may not follow the expected direction. That explains our two cues, one "weak" cue which is not always true, and one "exclusion" cue which is always true.

Dentals No comment here, except that the dental locus seems to be the more resistant one. We have thus specified a locus, that is a frequency region, and not only a direction as for the other cues.

Palatovelars Cues vary according to the vocalic context. We do not propose any cue if the adjacent vowel is a back vowel. If the vowel is central, we propose two kinds of cue. The "weak" cue is the well-known "velar pinch". This phenomenon may not be observed in cases of great coarticulation, revealed by the proximity of F2 and F3 at or near the center of a central vowel whose formants are normally spaced [7]. This proximity is characteristic of a palatovelar context, and we propose a "strong" cue to exploit this knowledge (this last cue is still on trial).

2.1.2 Acoustic cues provided by the burst in a (V1)CV2 context

V1 is not taken into account to build the following cues. For all the cues the frequency regions investigated are ranged, as an example it equals to [400 - 3,600Hz] for the palatovelar cues. The energy level values are adjusted according to the class of the following vowel.

Weak cue

Labials the energy maximum is situated in low frequencies.

Dentals the energy maximum is situated in high frequencies. A maximum situated in the locus frequency region is not taken into account [1].

Palatovelars the energy maximum is situated in front of F2 of back vowels, close to F3 of central vowels... [5]

Strong preference cue

Dentals the energy maximum is prominent in high frequencies [central and back vowels].

Palatovelars the energy maximum is very prominent, see before for its frequency value.

The two preceding cues are carefully built in order not to trigger off false alarms.

Strong exclusion cue

Palatovelars no energy in the appropriate frequency regions (where a peak should be detected).

Comments:
We do not think there can be any "strong" evidence of labial burst so no "strong" cue is proposed for (/p,b/). We are not so sure that the diffuseness can only be observed for labials and dentals, -some bursts of non-accentuated /g/ may also look rather diffuse-, so the diffuse characteristics are only used in "weak" cues, when necessary, as an example in a back context, to favour the labial candidate versus the palatovelar one. Note that the compactness is used in "strong" cue, as a characteristic of palatovelars.

2.2 Acoustic cue detection

Our approach assumes that the acoustic cue extraction satisfies the following requirements:

- a "strong" cue detector must not set off false alarms. It must be robust, not very sensitive and its implementation must rely on coarse burst characteristics;
- if a "weak" cue exists, corresponding to a preference one, its detector must be more sensitive to cover the cases where the acoustic cue exists but is not pronounced enough to be detected by the "strong" cue detector.

In the following, we describe the implementation of the acoustic detectors regarding the burst for the three voiceless stops (/p/,/t/,/k/) followed by back and central vowels. They have been implemented in Snorri [6].

In general, the burst release provides the most indicative information to identify the place of articulation. Therefore, our first aim was to separate the burst in two parts: the release and the friction noise.

2.2.1 Burst decomposition

In general, it is possible to distinguish on burst spectrograms two parts:

- the burst release;
- the friction noise, whose characteristics are not considered.

In order to separate finely the burst release from the friction noise, the spectrogram has been computed with a Hamming 4 ms window and a 1 ms time increment. Our burst decomposition is based on the description shown in Fig. 1.

The two burst parts are modelled by the energy concentrations corresponding to the highest spectral peaks extracted from the mean spectrum of each part. The boundary between the release and friction noise corresponds to the instant where the models of the two parts optimize a resemblance criterion with the spectrogram of the burst to be analyzed. We have tested several types of criterion:

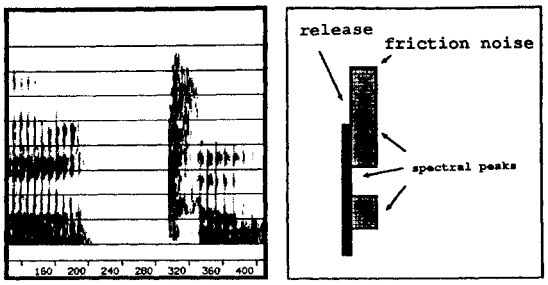

Figure 1: Spectrogram and burst decomposition

- correlation of the burst spectra with the computed models,
- spectrogram energy taken into account by the models,
- energy scattering in the models.

We have tested these criteria on a corpus of bursts, the results obtained allow us to retain for the decomposition the minimal scattering energy criterion; it presents the best trade-off between stability and localization error.

2.2.2 Burst release cues

We have imposed that the "strong" cue detectors return a value in {True, False }. A "weak" cue does not allow a direct stop identification, so its detector yields a value in the interval [0,1]. In this section, we describe only the palatovelar cue detectors.

2.2.3 Palatovelar detectors

/k/ followed by a back vowel In order to detect the palatovelar place of articulation, we have implemented two acoustic detectors:

- A "strong" cue called "strongBackVelarBurst", defined as a conjunction of the following acoustic events:
  - enhancement of a the spectrum between the frequencies 400 and 1,000 Hz (a intensive spectral peak), containing more than 50% of the energy above the spectrum mean;
  - this peak must be in the vicinity of the second formant;
- A "weak" cue called "BackVelarBurst" defined as a symmetrical aggregation of the following acoustic events:
  - a spectral peak located between 430 et 950 Hz,
  - the peak energy is twice more important than the energy of other spectral peaks.
  - the peak must be compact enough; the compactness has been defined by the peak inertia momentum related to the maximum frequency of the peak divided by the peak energy.

/k/ followed by a central vowel

- The "strong" cue "strongCentralVelarBurst" relies on the following acoustic events:
  - a prominent spectral peak between 1,000 and 2,500 Hz.
  - the peak must appear in the vicinity of second and third formant (at the voicing onset),
  - the compactness of this peak must be important (the diffuseness is less than 0.25).
- The "weak" cue "centralVelarBurst" relies on the following acoustic events:
  - a prominent spectral peak between the 1,300 and 2,000 Hz,
  - the compactness of this peak must be important (the diffuseness is less than 0.4).

2.3 Results

We have tested the acoustic cue detectors regarding the burst on a corpus of 561 voiceless stops (/p/,/t/,/k/) followed by back and central vowels. These bursts have been extracted from two BDSONS corpora: (i) a corpus of French monosyllabic words recorded within a carrier sentence, uttered by twelve male speakers and (ii) a corpus of continuous speech ("la bise et le soleil...") recorded by twelve male speakers. Table 1 and table 2 show the results we have obtained.

	number	labial	dental		palatovelar		activation rate	
		weak	strong	weak	strong	weak	strong	weak
/p/	81	72	0	3	0	6	-	89%
/t/	165	7	128	30	0	0	78%	18%
/k/	91	7	0	1	65	18	73%	20%

Table 1: Voiceless stops followed by back vowels

	number	labial	dental		palatovelar		activation rate	
		weak	strong	weak	strong	weak	strong	weak
/p/	56	52	0	1	0	3	-	93%
/t/	102	1	41	58	0	2	40%	57%
/k/	66	0	1	7	39	19	59%	29%

Table 2: Voiceless stops followed by central vowels

For the dental and palatovelar stops the "strong" cues are activated in 49% of cases, and give rise to only one false alarm. In 44% of cases, the "weak" cues lead to the correct stop and the error rate is 7%. These encouraging results confirm that energy concentrations and compactness of the burst release may be incorporated in a stop identification system (see part 3).

Moreover the absence of burst provides an important information for labial voiceless stops followed by central vowels.

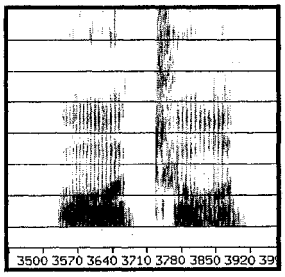

Figure 2: Spectrogram of /otu/.

3 Using the cues in a phonetic decoding system

3.1 Motivations

Our contention is that the recognition process must be consistent: identification is permitted only if the detected cues are mutually consistent. A contradiction appears when either a "strong preference" cue contradicts a "strong exclusion" cue or "strong exclusion" cues dismiss all solutions proposed by extracted "weak" cues. Otherwise, one of the previous step of the recognition process must be questioned. This step can either be the segmentation task or the extraction of cues performed by automatic acoustic detectors. Thus, we used ATMS[1]-like hypothetical reasoning techniques [4] to maintain the consistency of our deductions. In an ATMS, every fact is labelled with the hypotheses which lead to its deduction. If facts or hypotheses are inconsistent, they cannot be associated to make a deduction. Such a system has been developed in our research group [3] and we built up a stop identification system on this system (using cues provided by the burst and the formants). More details on our approach can be found in [2].

3.2 An example

Let us follow the identification task of the stop presented on the spectrogram on Fig. 2. The acoustic detectors provide the following cues:

- pronounced dental burst ("strong preference" cue for /t,d/) P
- no energy in front of F2 ("strong exclusion" cue for /k,g/) E

If a "strong preference" cue is found, "weak" cues are not taken into account. Thus, formant transitions cues are not considered in this example. Our method consists in explaining the extracted cues in assuming models corresponding to these cues. In our example, the cue P leads to the assumption of a dental. Furthermore, the recognition of a palatovelar is excluded because of the cue E. Hence the only solution found here is a dental.

[1] Assumption-based Truth Maintenance System

3.3 Results

We tested our system on a corpus of stops in back vowel context. Our corpus contains continuous speech and isolated monosyllabic words. The continuous speech part is made of 15 occurrences of stops spoken by male speakers. The other part has 15 occurrences, each of them spoken by two male speakers and a female speaker (without any adaptation). The results are:

- 46/60 of correct recognition (as the only possible solution)
- 53/60 of correct solutions in first or second position

The main failures are caused by an erroneous detection of formants (F2 is missed). Improvements on the formant tracker algorithm should be significant on our results.

4 Concluding remarks

Our identification strategy which is based on two levels of acoustic cue appears to be efficient inasmuch as our "strong preference" cue allows 50% immediate identification. Perceptual studies will be undertaken to investigate how the performances of a stop identification system based on acoustic cues resemble human performances for the same task. Such perceptual information will be particularly interesting especially for front vowels and more precisely for /i/, a context in which stop recognition is acknowledged as being difficult.

References

[1] S. E. Blumstein and K. N. Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustic Society of America*, 66(4):1001–1017, October 1979.

[2] A. Bonneau, F. Charpillet, S. Coste, J.-P. Haton, Y. Laprie, and P. Marquis. A Model for Hypothetical Reasoning applied to Speech Recognition. *Proc. of the 10th European Conference on Artificial Intelligence, Vienna*, 1992.

[3] F. Charpillet, Ph. Théret, and J.P. Haton. X-TRA : un moteur d'inférence comportant deux modes de compilation des règles TREAT et RETE et un système de maintien de vérité de type ATMS. *Actes des journées internationales des systèmes experts, Avignon*, 1988.

[4] J. de Kleer. An Assumption-Based Truth Maintenance System. *Artificial Intelligence*, 28(2):127–162, 1986.

[5] E. Fisher-Jorgensen. Acoustic Analysis of Stop Consonants. *Miscellana Phonetica*, II:42–59, 1954.

[6] D. Fohr and Y. Laprie. Snorri: An Interactive Tool for Speech Analysis. *Proc. of EUROSPEECH, Paris*, pages 669–672, 1989.

[7] J. Vaissière. Effect of Phonetic context and timing on the f-pattern of the vowels in continuous speech. *Proc. of the 11th International Congress of Phonetic Siences*, 5:43–46, 1987. Tallin, Estonia.