Elitist identification of stops from formant transitions

Anne Bonneau, Yves Laprie

LORIA/CNRS and INRIA

Vandoeuvre-lès-Nancy, France

E-mail: anne.bonneau@loria.fr, yves.laprie@loria.fr

ABSTRACT

We present a set of strong cues for stop place of articulation, based upon formant transition slopes. This set complements a first one defined from stop bursts. Strong cues, which identify or eliminate a phonetic feature with certainty (no error is allowed), are not systematically fired. They can be exploited in language learning (cue enhancement) or in ASR to reduce the search space during the lexical access. The detection of the formant transitions and the cues is entirely automatic. To that purpose, we elaborate a robust formant tracking detector able to follow the transition trajectory. We obtained an average firing rate of 30 % on a clean corpus made up of sentences.

1. INTRODUCTION

In previous works [1,2], we have shown that there exist "strong" cues, as we call them, which can identify or eliminate a phonetic feature with certainty (no error is allowed). Such a decision is possible in few cases, when an acoustic cue is well marked and has a high power of discrimination. During strong cue detection, we must fulfil two requirements: to make no error on the one hand, and to obtain a relatively high firing rate, on the other hand. The notion of strong cue must not be merged into that of « robust » cue or landmark which are systematically fired and are allowed to make some errors. Strong cues can be exploited either in language learning in order to enhance the most reliable cues or in ASR to provide "confidence islands" and reduce the search space during the lexical access.

The transition cues complement a first set of "strong" cues already defined from stop bursts, which exploits the frequency and the compactness of the most prominent peak. This set was made up of positive and negative cues, which respectively identify or rule out a phonetic feature with certainty. Whether formant transitions, which vary a great deal with vocalic context and speech rate, could provide highly reliable cues for stop place in continuous speech, is far from evident. Indeed, D. Kewley-Port [3] has shown that F2 and F3 onset frequencies from stop-V isolated syllables uttered by one speaker provide only context-dependent cues for stop place. Moreover, she noted that the frequency boundaries of the (contextual) regions coding distinctive stop categories were very close together so that, at a perceptual level, "the distinctions between stop categories are likely to be lost". We add that it is also highly probable that they won't resist to speaker or speech rate variations. In his study concerning V-stop-V logatoms, Öhman [4] has shown that VC and CV transitions are influenced by the vowel situated on the other side of the stop (the transconsonantal vowel). In some cases, the transition slopes of identical CV (or VC) sequences with a different transconsonantal vowel were even radically opposed. Although Lindlöm [5], reconsidering Öhman's data (obtained from one speaker), found specific regions characteristic of each stop category in the F2-F3 acoustic space, these regions were once more very close together and would probably not resist to variations.

If we want to find very reliable cues in continuous speech automatically, we have to face two main problems:

- finding cues which are "resistant" to different speakers, contexts, speech rates and to the effects of non-surrounding sounds (such as the phoneme P in stop-V-P sequences or P-V-stop sequences),

- detecting these cues automatically, although formant tracking is especially difficult on highly variable segments such as the frontiers between stops and vowels.

To cope with this difficult task, we took only F2 into account, a formant which is both more informative for stop place identification and (often) more easy to detect than F3; and we considered only transition slopes. We also defined only negative cues, which rule out a consonant class.

Our method was the following: we defined a first set of cues available in #CV context, then we examined the effect of the transconsonantal vowel in V-stop-V sequences on this first set and restricted the application of some cues when necessary. Finally, we determined how this set of cues could be applied in larger contexts. We will describe only CV transition slopes, since, at least for dentals and labials, we can consider that VC slopes are just the opposite.

All the transition slopes are detected automatically, so we will describe the automatic formant tracking algorithm,

ISBN 1-876346-48-5

then we will present a first set of cues for unvoiced stops followed by back and central vowels.

2. FORMANT TRACKING

2.1. The algorithm

Most of the formant tracking algorithms uses dynamic programming to connect spectral peaks (LPC roots, cepstrally smoothed spectra peaks …) from one spectrum to the following. However, formant tracks resulting from these algorithms cannot be used directly because they are not sufficiently smoothed and a complementary smoothing might debase the relevancy of results. Therefore, we accepted a method that combines intrinsically tracking and smoothing. In [6], we proposed an algorithm based on a regularization technique used in computer vision called "active contours" or "snakes". The objective is to find a curve such that it is as close to the spectral maxima (i.e. formants) as possible, yet, on the other hand as smooth as possible. This gives rise to an iterative process called active contours.

Under the influence of the spectrogram energy the iteration process generates a sequence of curves which converges to the nearest formant trajectory. As with most iteration methods an initial solution must be provided. We therefore developed an interpretation algorithm which extracts lines of spectral peaks and label them in terms of formants by using knowledge of the frequency domain and energy level of formants. The quality of final formant trajectories mainly depends on the quality of the initial solution which is sometimes not relevant. Sun proposed a less sophisticated method [7] based on natural cubic splines, the main quality of which is to use a very simple initial solution for a formant which is the average value of LPC peaks in a frequency domain defined according to the formant. The second characteristic of Sun's method is to use a center-of-gravity method which corresponds to the calculation of the probability of a spectral frequency to belong to a given formant. This prevents two formants from getting too close together.

Our new algorithm exploits the two methods presented above in the following way. We choose straight lines as initial trajectories. For each formant we combine the influence of the spectrogram and that of the probability of a frequency to belong to this formant, we compute one iteration of the active curve and re-estimate the frequency-to-formant affiliation probability. We iterate this process until the trajectories stabilize. Unlike our previous algorithm, trajectories are computed simultaneously. Therefore, the deformation of one trajectory influences the others, which substantially improves the global goodness of fit between formants and spectrogram.

Figures 1 and 2 illustrate the interest of the tracking based on concurrent curves. Fig. 2 shows the result obtained without correction with non concurrent curves. In particular, one notes that there are gross errors near 300 ms and 500 ms due to the existence of a nasal segment between 240 and 300 ms. This segment made the interpretation in terms of oral formants to fail and this error propagates to the neighbour oral vowel (between 300 and 360 ms). The concurrent tracking (Fig.1), on the other hand, gives quite good results even if the initial trajectories obtained by averaging LPC roots values are very rough. The competition between formant curves allows the energy of the spectrogram to be captured at the best.

2.2 Spectral analysis for formant tracking

We pay a particular attention to the calculation of the spectrogram which is used for formant tracking. There are basically two methods: linear prediction and cepstral smoothing.

Formants are often extracted from a spectrum which corresponds to linear prediction. Weaknesses of this method are well known: formants are "attracted" by harmonics and non vowel sounds are not well analyzed. Beside these difficulties we observe that the amplitudes of formants can vary considerably from one frame to the next one. Errors on amplitude are often about 10 dB and can give rise to major problems when amplitude is required (copy synthesis for instance).

Cepstral smoothing presents less weaknesses (nasal sounds are analyzed properly, for instance) but is sensitive to the position of the analyzing window in the signal. Furthermore the level of the resulting spectrum is below harmonic peaks and close formants are often merged. Thus we accepted the "True envelope" which deforms the smoothed spectrum so that is approximates harmonics correctly (Figure 3).

This method derives from the cepstral smoothing and has been originally proposed by Imai and Abe [8].

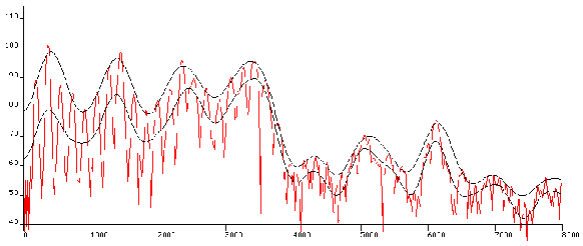

Figure 3. Narrow band spectrum, cepstrally smoothed spectrum (the smoothed curve at the bottom of the figure) and "true envelope" (close to harmonics).

The principle is to use the cepstral smoothing and to correct it iteratively by reducing the contribution of spectral values above the smoothed spectrum. These values represent harmonics and give rise to a correction term to the cepstral coefficients. Although this method is somewhat time consuming as two Fourier transforms are needed at each iteration, it gives excellent results on real speech signals. Note that this method gives results close

ISBN 1-876346-48-5

to those obtained by discrete cepstra proposed by Gallas [9]. Indeed, these two methods use cosine functions. The advantage of the "true envelope" method is that no prior peak picking is required.

3. STRONG CUES FROM FORMANT TRANSITIONS

3.1. #CV context.

Labial transitions (#CV). The slopes of the F2 and F3 transitions between labials and non-rounded vowels are rising, labialisation lowering the initial formant frequencies of a following non-labialised vowel. For rounded vowels, these slopes are either rising or slightly falling, but never strongly falling.

Dental transitions (#CV). Dentals are the only consonants which have a relatively stable locus (around 1600 Hertz). As a consequence, the slope of F2 transition depends upon the F2 frequency of the vowel. It is very well marked (strongly falling) before back vowels, whose F2 is low.

Palatovelar transitions (#CV). There is no systematic pattern for palatovelar transitions before back and high front vowels, even if they tend to be rising in the first context, and relatively flat in the second one. Before central context however, the palatovelar transitions are well marked, especially F2 transition which is falling, and constitutes with the rising F3 transition "the velar pinch".

If we consider the transition slopes for each vocalic context separately, we observe that some patterns never occur for one place, but sometimes appear for at least one other place. These oppositions enable us to define a set of strong (exclusion) cues available in CV context (Fig.4).

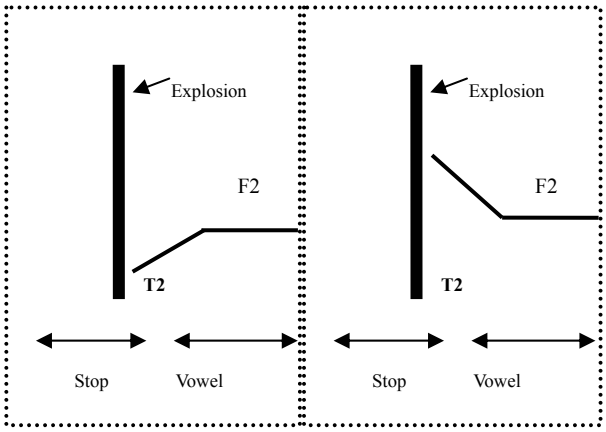

Figure 4. Representation of the exclusion cues :the figure on the left represents the labial exclusion cue, the figure on the right represents the dental exclusion in back vocalic context, and the palatovelar cue in central contexts.

Back context (#CV). Labial and dental well marked slopes suggest the two following exclusion cues. Dentals are excluded when the F2 transition is rising. Labials are excluded when the F2 transition is strongly falling.

Central context (#CV). Two exclusion cues are derived from labial and velar transition slopes which are both stable and in opposite direction.

Labials are excluded when the F2 transition is falling. Palatovelars are excluded when the F2 transition is rising.

3.2. Transitions in VCV sequences.

Is the set of cues defined in the previous chapter still available in continuous speech? We will first consider intervocalic (VCV) contexts, and try to determine the effect of the transconsonantal vowel, pointed out by Öhman, on the well marked transition slopes chosen to define the cues.

Labial consonants. During labial articulation the tongue body is free to move from the articulation of the first vowel to that of the second. If the first vowel of a VCV sequence is more acute than the second one, and if the vowels are coarticulated, the CV transition will be falling instead of being rising, which is in contradiction with our labial exclusion cue. Even if such intervocalic coarticulation is not frequent, at least for unvoiced stops, the application of the labial exclusion has to be restricted to avoid any false alarm. In order not to limit too much the application of the cue, we also verify that the transition of the transconsonantal vowel is compatible with this intervocalic coarticulation (i.e. is not rising for VC). The following rule is intended to eliminate possible coarticulatory effects. "If the transconsonantal vowel is more acute than the vowel under consideration, and if its F2 transition is compatible with an intervocalic coarticulation, the labial exclusion cue is not applied". Practically, we consider than a vowel is more acute than the other if their F2 frequency at the frontier with the stop differ from more than 50 Hertz.

Palatovelar consonants. We know that the tongue body (the main articulator for palatovelars) anticipates the position of the following vowel during the production of the consonant, but we don't know the extent to which the transconsonantal vowel affects the position of the tongue body at the instant of release, and, as a consequence, the CV transition slope. If the transconsonantal vowel is back (and rounded), its influence could change the direction of the CV slope in central context and invalidate the palatovelar exclusion cue. A rule similar to the one proposed for labials could eliminate such possible coarticulation.

Dentals. We considered that the transconsonantal vowel has no important effect on the dental place of articulation, and doesn't put into question our dental exclusion cue (back vowel context).

ISBN 1-876346-48-5

3.3 Continuous speech

Sometimes the transition between the vowel and the following phoneme begins at vowel onset during what should be the stop-vowel transition, and put into question the set of strong cues. This happens when the vowel is very short. We use this brevity to eliminate such cases.

4. RESULTS AND CONCLUSION

We have tested our set of strong cues on two corpora made up of sentences; the first one was recorded in laboratory condition by four male speakers and the second one, noisier, in an office by 16 male speakers. The segmentation was automatic but verified by hand.

From the first corpus (500 stops) we obtained an average firing rate in 25 % of the cases, with no error. We actually test the cues on the second corpus (made up of 1700 stops). First results show that the average firing rate will be slightly lower. The transition cues will be combined with the strong cues designed from stop bursts. The whole set will be exploited for ASR and in Language Learning.

REFERENCES

[1] A. Bonneau, S. Coste, Y. Laprie « Two level acoustic cues for consistent stop identification", Proceedings of the *International Conference on Spoken Language Processing*, Banff (Canada), 1992.

[2] A. Bonneau, S. Coste, Y. Laprie, "Strong cues to identify features with certainty", *International Congress of Phonetic Sciences*, Stockholm., 1995.

[3] D. Kewley-Port, "Measurements of formant transitions in naturally produced stop consonant-vowel syllables", *JASA*, vol. 72.2, pp. 379-389, 1982.

[4] S.E.G. Öhman "Coarticulation in VCV utterances: spectrographic measurements", *JASA*, vol. 39, pp. 151-168, 1966.

[5] B. Lindblöm "Role of articulation in speech perception: clues from production", *JASA*, vol 99, pp. 1683-1692, 1996.

[6] Y. Laprie and M.O. Berger "Cooperation and regularization of speech heuristics to control automatic formant tracking", *Speech Communication*, vol. 19. pp. 255-270,1996.

[7] D.X. Sun "Robust estimation of spectral center-of-gravity trajectories using mixture-spline models" *European Conference on Speech Communication and Technology,* vol. 1, pp. 749-752. Madrid, 1995.

[8] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepctral method", *Trans IECE*, J62-A, 4. 1979.

[9] T. Gallas and X. Rodet, " Generalized functional approximations for source-filter modelling" *European Conference on Speech Technology,* Genova, 1991.

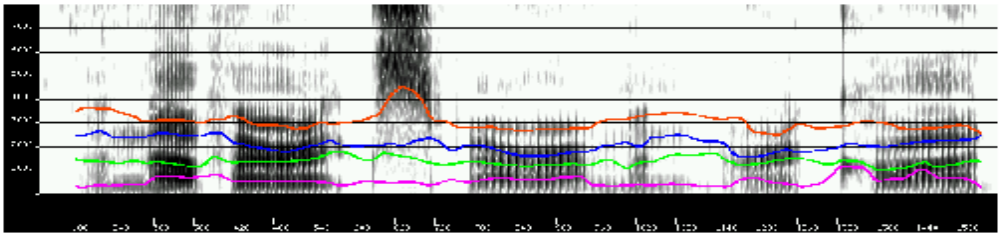

Figure 1. Tracking with competitive curves (60 iterations)

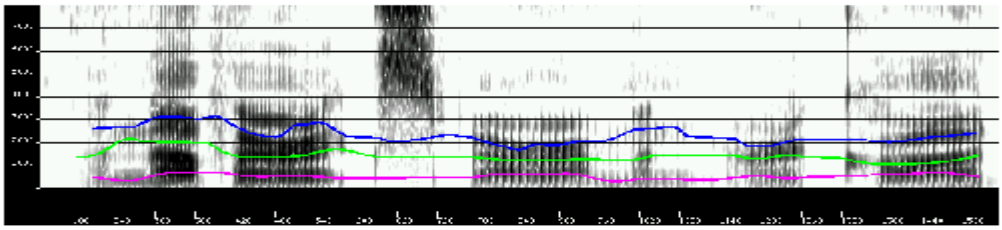

Figure 2. Tracking without competitive curves.