Busrt segmentation and evaluation of acoustic cues

Yves Laprie and Anne Bonneau

Speech Group
LORIA
Vandœuvre-lès-Nancy, France
Yves.Laprie@loria.fr, Anne.Bonneau@loria.fr

Abstract

This paper investigates burst segmentation for the evaluation of acoustic cues used to identify unvoiced French stops. Unlike other works which utilize a fixed length window, our approach consists in segmenting bursts into transient and frication noise. The transient is found by minimizing the sum of spectral variances of transient and frication noise over the burst. The spectral variance criterion has the advantage of being sensitive both to energy deviations and spectral variations. Additional correction procedures augment the robustness of the segmentation method, then we compare the bursts characteristics extracted from bursts segmented or not. The discussion of the results according to the stop place of articulation concludes the paper.

1. Introduction

This paper takes place in the definition of acoustic cues for identifying French unvoiced stops in vocalic context. Our work focuses on the elaboration of cues with a high power of phonetic discrimination and well marked from an acoustic point of view. They can be positive - in this case they lead to an almost sure identification of sounds -, or negative - in this second case they rule out the identification of sounds -. These cues exploit formant trajectories and spectral or energetic characteristics of bursts.

Burst of stops can be decomposed into three segments [3], transient, aspiration (absent in French) and frication noise. Formant transitions can be seen in the fricative noise and the aspiration. A burst begins at the release of the stop constriction located at the place of articulation, and ends with the first vocalic period (which can be easily found in a unvoiced stop-vowel sequence). Unlike the noise of fricatives, burst of stops does not correspond to a sustained articulation but to the transition from one place of articulation to the following (that of a vowel, generally). Consequently, the burst spectrum changes rapidly over time and, in most cases, the spectral cues are most discriminating at the beginning of the burst than at its end. Indeed, the closer to the subsequent vowel the spectrum is, the more it is dominated by peaks corresponding to the formants of the vowel. This was clearly visible in the spectra exhibited by Krull for voiced stops [4], performed at two points in time (t1 : from 0 to 10 ms after the consonant release and t2: from 10 to 20 ms after it).

It seems therefore judicious to evaluate cues for the stop place of articulation just after the release of the consonant articulation. However, the segmentation of the burst is a difficult issue and, generally, only a fixed length segment at the beginning of the burst is taken into account (26 ms in the work of Stevens and Blumstein [1], or 10 to 15 ms in the work of Zue [6]). In order to achieve a very precise evaluation of acoustic cues on burst, we have devised a method to segment bursts into two or three segments.

In the next section we will present the segmentation method, then we compare the bursts characteristics extracted from bursts segmented or not. The discussion of the results according to the stop place of articulation concludes the paper.

2. Burst segmentation

2.1. How to decompose the burst?

As we are working on French, the decomposition of the burst in two parts - transient and frication noise - is sufficient.

However, in the case of multiple transients it is necessary to add an extra part which represents the first transient among the multiple transients, when it is not the most prominent. In this case, there are three parts exceptionally: (i) the first transient followed by a silence, (ii) the next transients which are merged together to form the main transient, (iii) the frication noise.

This decomposition in three parts is necessary when the silence between the first and the next transients cannot be negligible, i.e. approximately 10 ms.

The first step is to locate the burst onset and to measure its duration. As we are working on the CV context, the end of the burst corresponds to the voicing onset as it is given by the fundamental determination algorithm we developed previously [5]. Using elaborated correction procedures allows a very precise voicing onset to be obtained (with at most a 8 ms error which is the time shift during the F0 determination).

All the burst localization as well as decomposition computations are performed from a large band spectrogram with a 4 ms window and a 1 ms time shift. These choices thus allow a temporal precision of 1 ms.

The burst onset corresponds to a significant apparition of energy. The energy threshold used to locate the onset is rather weak but we impose that a substantial peak appears within 10 ms after the onset. This peak in energy must be at least 6 dB higher than the silence level set to display the large band spectrogram.

The first transient of a multiple transient is searched for in the form of a strong and brutal peak followed by a short silence and then a peak higher than the first one. We impose that the first transient is not more intensive than the main transient because we calculate acoustic cues from the characteristics of the
main transient. In the case where the first transient is more intense it plays the role of the main transient and the following multiple transients are merged with it. From a practical point of view, the first transient is often weaker than the others. That is consistent with the choice we did. Most of the time, multiple transients are observed for /k/ and sometimes for /t/ in the context of a front vowel.

2.2. Which criterion to segment a burst?

The observation of a very large number of bursts leads us to distinguish several main classes of bursts, which can be useful to guide the development of a segmentation algorithm. For each of these classes we give a schematic representation (see Fig. 1), the profile of the energy curve and an example.

The most promising two classes with the aim of using acoustic cues of a high power of phonetic discrimination are the first ones. This first class corresponds to bursts that present a very strong transient dominating what remains of the burst. Moreover, the spectra of transient and frication noise are clearly different. For an average burst, spectra of transient and frication noise differ but the transient energy is not prominent compared against that of the frication noise. In the third case the transient is not very informative because of its weak energy. The segmentation is also more difficult (and thus less accurate) and presents less interest as the transient cannot be used to identify the stop with certainty.

Two methods can be imagined to segment bursts. The first relies on a energy criterion to locate strong transients, and on a spectral criterion to segment transients when there are not very prominent. The second method appeals to a single criterion for the aspects of energy and spectrum. The criterion accepted must be very robust so that this method can be efficient. We studied two criteria with this aim in view.

Let be \(X(e^{j\omega})\) the Fourier transform of the speech signal and \(S(e^{j\omega}) = \max(20\log_{10}|X(e^{j\omega})| - dB, 0)\), i.e. the energy above the threshold of silence in the spectrogram computation.

The first criterion is the sum of correlation between the spectra of a speech segment and their average:

\[
\text{Corr}(t_0, t_1) = \sum_{t=t_0}^{t_1-1} \frac{\sum_{\ell=t}^{t_1-1} S(t_0, t_1) S(e^{j\omega}) \overline{S(e^{j\omega})}}{\sqrt{\sum_{\ell=t}^{t_1-1} S(t_0, t_1)^2} \sqrt{\sum_{\ell=t}^{t_1-1} S(e^{j\omega})^2}}
\]

where \(K_0\) and \(K_1\) represent the spectral range to be taken into account. The criterion to be maximized is thus:

\[
\text{Corr}(t_0, \text{limit}) + \text{Corr}(\text{limit}, t_f)\]

where \(t_0\) and \(t_f\) are the times of beginning and end of the burst, and \(\text{limit}\) the boundary between the transient and the frication noise.

The second criterion is the spectral variance calculated on the speech segment:

\[
\text{VarSpec}(t_0, t_1) = \frac{1}{t_1 - t_0} \sum_{t=t_0}^{t_1-1} \sum_{\ell=t_0}^{t_1-1} (S(t_0, t_1) - \overline{S(e^{j\omega})})^2
\]

and the sum of spectral variances of the two parts has to be minimized over the whole burst.

The correlation criterion is especially sensitive to the spectral homogeneity as the influence of energy is removed by the normalization. Clearly, the variance criterion is more sensitive to energy as the deviation from the average energy is taken into account.

Preliminary experiments showed that the variance criterion corresponds well to our objective. It gives good results in the cases where the transient is a strong energy peak or presents a spectrum far from that of the frication noise. We thus accepted the spectral variance criterion. Consequently, as this criterion can be applied with benefits in most of the bursts the segmentation method we devised exploits one criterion only.

2.3. Augmenting the segmentation robustness

Segmentation errors originiate either in the acoustic structure of the burst itself, or in the context the segmentation is used in. This section describes the methods used to augment the robustness of the segmentation against errors made on the determination of the burst onset or the voicing onset. Indeed, this algorithm relies on the assumption that the burst can be decomposed into two parts. It does errors when this assumption is not satisfied, because either the voicing onset is located to late, or a spurious noise comes before the burst.

The first difficulty is linked with the inaccuracy of the voicing onset determination. Indeed, taking into account one or several periods of voicing may move the transient boundary. In fact, the boundary may be found almost everywhere between the true boundary and the burst end which erroneously falls into the vowel. To eliminate this kind of error the influence of the energy at burst end must be reduced. For that purpose we replace the deviation term in the spectral variance with:

\[
\left(\frac{S_i(e^{j\omega}) - \overline{S(e^{j\omega})}}{\cos^2\left(\frac{\omega}{2\pi}\right)}\right)\]

where \(t_i\) and \(t_f\) are the burst boundaries. That means that:

\[
t_i < t_0 < t_f < t_f.
\]

The exponent \(\alpha\) is used to limit the influence of the last spectra in the burst and depends on the accuracy of the voicing determination. The reduction of energy influence at the burst end enhances the strength of energy at the burst onset and could leads to too early a boundary. In practice, energy at the burst onset is almost always clearly stronger than that at the burst end and the boundary is therefore not modified.

The second difficulty is linked with the existence of spurious noises during the closure, which give rise to false alarms. These noises are detected because the energy threshold for detecting burst onsets is weak so that bursts without a strong transient can be located (cf. the third case of Fig. 1).

These false alarms are characterized by a small noise, then a silence and the true burst. Several small spurious noises may come before the burst but that does not modify the correction procedure described below.

The spurious noise always corresponds to the transient, the silence may be merged with either the transient or the frication noise. The correction of these two situations is described here:

- **the spurious noise and the silence form the transient** The average spectrum of the transient is evaluated after the segmentation. If the silence is merged with the spurious noise then the average spectrum is very weak or even null; it thus cannot represent that of the transient. The segmentation procedure is iterated by shifting the beginning of the burst until the average spectrum becomes at least 3 dB higher than the threshold of the silence in the spectrogram computation.

- **the spurious noise plays the role of the transient** This means that the silence and the true burst are merged together and that they represent the frication noise of the spectrum. The frication noise is segmented after the first segmentation. If there is a silence with a significant duration that means that the frication noise is in fact
The acoustic cues are all the more effective as the prominence is sufficiently high provided that the frequency of the peak is the expected one for a stop place. We thus projected the consonants in a plane formed by the frequency of the main peak (x-axis) and its prominence (y-axis). We tested two criteria to characterize the prominence of the peak: (i) its relative energy with respect to the average energy of the spectrum, and (ii) the difference in energy between this peak and the second most prominent peak. The second criterion appeared to be slightly more effective and we adopted it. The compactness and diffuseness of the spectra were also evaluated, but they were less effective (and their definition is not trivial) so we have not taken them into consideration here, although they do have a role to play in stop place identification. We can nevertheless observe that the high prominence of the main peak of most velars bursts is related to its compactness.

In order to validate our method, we used the characteristics of the main peak (its frequency and prominence) to compare the discrimination power of the whole burst spectrum (its energy was averaged over the time) with that of the burst onset, segmented with our method. The whole burst was segmented from the burst release (determined by our method to localize burst) to the first period of the following vowel. We tested the procedure on a limited set of /p?-stop-V/ sequences extracted from two French corpora. The first corpus (a subset of BDSONS) contained isolated words spoken by 5 male speakers. The second corpus contained 22 read sentences made up of stops and vowels and repeated 3 times by 4 male speakers. We analyzed five hundred consonants approximately. We will comment the results for two vocalic contexts in detail, the front rounded and the back context.

3. Segmentation and evaluation of burst characteristics

Rather than inspecting segmentations produced by our method visually, it seems more relevant to investigate whether burst cues are favourably influenced by our segmentation.

The main acoustic cues provided by the burst spectrum for the identification of stop place are the frequency of the most prominent peak (designated below as the “main peak”) and the spreading of energy (its compactness or diffuseness). Stevens and Blumstein[1] defined the alveolar spectra as diffuse/rising, the labial spectra as diffuse/falling and the velar spectra as compacts. Most of the times, the frequency of the main peak is higher than 2200-2500 Hz for dentals (excepted when the peak corresponding to the dental locus dominates the spectrum), lower than this frequency region for labials, and corresponds to the frequency of the F2 (before back and front rounded vowels) or to the F3 (before unrounded front vowels) of the following vowel for velars. The intra- and inter-contextual variability of this cue for each stop place is very important (see Fig. 2). Consequently, the frequency of the most prominent peak cannot be a reliable cue if the main peak is not sufficiently prominent (for the labials, the burst is generally too weak to allow a reliable identification of this consonant[6, 2]).

Figure 1: The three classes of burst. Black rectangles in the schematic representation of burst (top) correspond to spectral prominences. For each example (from left to right /ka/, /kœ/, /ty/) the spectrogram (one graduation each kHz) and the energy profile (one graduation each 5 dB) are given. Our automatic segmentation is superimposed onto the spectrograms.

made up of a silence followed by the true burst. If the duration of the silence exceeds 20 ms one considers that the frication noise actually corresponds to the burst, and that it has to be segmented.

This double correction on the transient and the frication noise eliminate almost all the errors due to spurious noises.

It should be noted that these corrections do not eliminate multiple bursts which correspond to peaks of energy which are significantly higher and followed by a rather short (generally less than 10 ms) silence.
4. Discussion and concluding remarks

In the following discussion we will compare the results obtained for the whole burst with those obtained for the burst onset segmented according to our method. For front rounded vowels, results clearly showed that the consonants were more concentrated in the F2-F3 region of the following vowel (approximately 1700-2400 Hz) when the cues were calculated from the whole burst. This concentration was due to the lowering of the most prominent peak of all labials whose frequency was higher than 2500 Hz and of most dental consonants. We also observe the drastic decrease of energy of the main peak for dentals, so that, even the dentals with a high frequency peak could not be identified with a high level of certainty. Finally, the increase of energy for some labials, due to an intense frication noise, combined with the decrease of energy for velars, reduced the number of clear evidences of this last class of consonants. The same great tendencies can be observed for the back context. The lowering of the frequency of the main peak for some dentals and the greater concentration of labials in the F2 frequency region of the following vowel, i.e. the region of the main peak for velars, reduced the discriminating power of the frequency as a cue to stop place when calculated from the whole burst. This reduction was amplified by the (drastic) decrease in the energy of the main peak of dentals, and that of velars (both absolutely and relatively to labial peaks).

As expected, these examples showed that the burst onset, conveniently segmented by our method, allowed a better discrimination between the three places of consonants than the full burst. We obtained the same kind of result, but less evident, for the central and the front unrounded context.

5. References