

AUTOMATIC SEGMENTATION OF SPEECH INTO SYLLABLES.

P. Mertens.+

ABSTRACT

A multiple pass procedure for the automatic segmentation of syllabic units is described which involves (1) a broad segmentation triggered by the dips in the intensity curve of band-pass filtered speech, (2) a further segmentation on the basis of the shape of the curve, and (3) the readjustment of the syllabic nucleus within syllable boundaries, based on the intensity of the unfiltered speech.

INTRODUCTION

The segmentation procedure has been designed as a part of a system for automatic recognition of intonation patterns in French. In intonation analysis, a correct identification of the syllabic nucleus is required, even at fast speech rates.

Syllabification strategies based upon intensity measurements fall into three types.

The first type (ref 1,2,8) defines the syllable as the segment between the points on the time axis, before and after a local maximum in the intensity curve of the speech signal, that are at a certain threshold level below the peak, provided the segment has some minimum length. The latter condition avoids false alarms from plosives being detected as syllables. The nature of the intensity measurement (peak to peak amplitude, peak amplitude, mean amplitude, rms intensity) has no significant effect on the results.

The major problem with this method is the frequent lack of segmentation in the context V C V where C is a voiced consonant, in particular a nasal or a lateral.

Compared with vowels, voiced consonants show a concentration of energy in the spectral region below 500 Hz. If this region is filtered out from the speech signal, voiced consonants show up as dips in the intensity curve. This effect can be achieved in various ways: by high-pass or band-pass filtering (ref 3), by frequency normalization of the power spectra (ref 4), or by loudness measurements (ref 6,7,11). In this second type of segmentation, syllables are detected from the resulting intensity curve in the same manner as for the first method or, alternatively, using Mermelstein's convex hull method (ref 4).

The filtering has some negative effects on the accuracy of the syllabic nucleus boundaries. First, overall vowel intensity is affected. In closed vowels, part of the first formant is filtered out due to the lower cut-off frequency of the filter. When these vowels stand next to inherently loud consonants, such as [l] and [R], the syllabic nucleus is shifted towards the consonant. Second, syllabic nuclei are generally shortened. This effect increases as the upper cut-off frequency is lowered. The higher formants of the vowel, which in stressed vowels exhibit a progressive energy drop during vowel emission, are attenuated. As a result the right side of the syllabic nucleus is shifted to the left and the last part of the pitch contour is excluded from the nucleus.

In order to minimize the drawbacks of the filtering method, without sacrificing its advantages, the two previous methods can be combined (type 3) in such a way that the first segmentation is modified by a parallel

+ University of Leuven, Blijde-Inkomststraat 21, B-3000 Leuven, Belgium.

segmentation based on the intensity curve of the filtered speech (ref 5,9,10).

However, the initial segmentation remains decisive for the overall performance of the detection algorithm, since the parallel segmentation will correct the initial one only within the boundaries of the initial syllable.

In a preliminary experiment, the three methods have been implemented and extensively tested on continuous speech data for French. Remaining errors must be attributed to the presence of [l] and [R], to the absence of a dip in the intensity curve, or to the presence of a dip within the vowel.

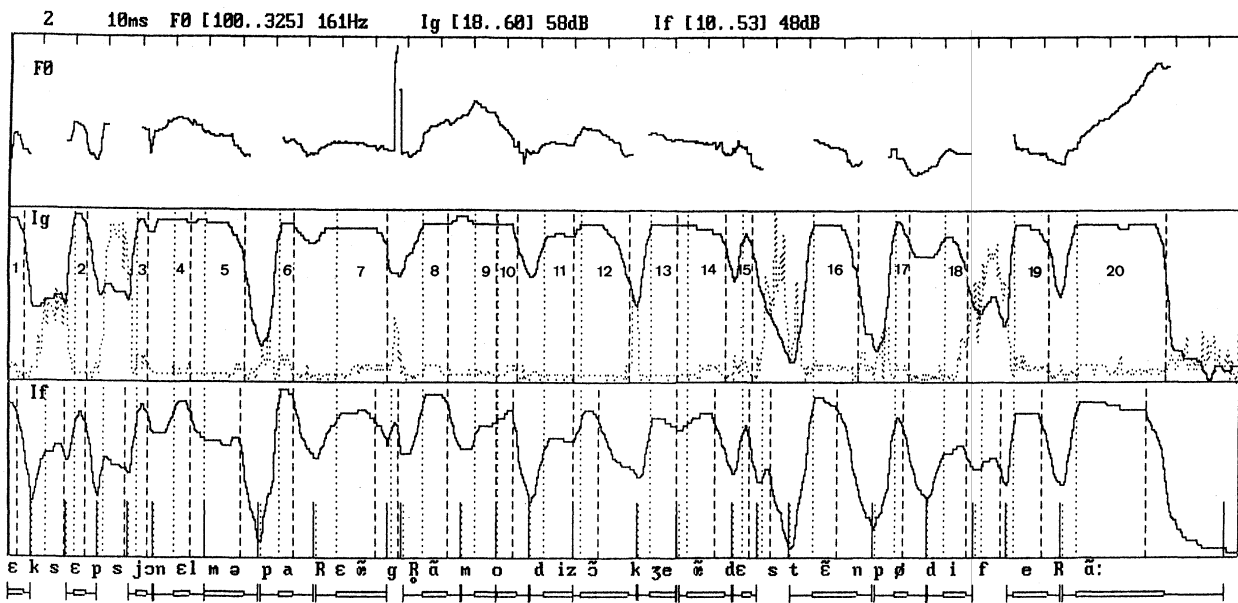
DESCRIPTION OF THE ALGORITHM

The segmentation procedure follows a direction opposite to that of the combined method (type 3). In the combined method, the segmentation obtained by the analysis of local dips in the intensity curve (I_g) is corrected by the inspection of the intensity curve of the filtered signal (I_f). In our program, the initial segmentation is based on the evaluation of dips in I_f (step 1) as well as on the analysis of the form of the curve within tentative syllabic boundaries (step 2); the syllabic nucleus is adjusted within syllable boundaries using a variable threshold which is applied to variations of I_g ; false alarms are avoided by checking several acoustic properties (step 3).

The speech signal, low-pass filtered at 4kHz (anti-aliasing), is sampled at 8kHz. Every 5 ms, intensity I_g is measured as the rms intensity of the samples in a 25 ms window. The intensity data are converted to a logarithmic scale and smoothed (moving average of 3 values). The intensity of the filtered signal, I_f , is obtained in the same way except for an additional (digital) band-pass filtering (500 - 3000 Hz, 24 dB/oct). The zero-crossing rate is computed on the samples in a 5 ms window. An analog pitch detector is used for fundamental frequency measurements. For I_g and I_f the absolute level threshold (cf. infra) is defined as 40 % of the dynamic range for the speech segment (5s) analysed.

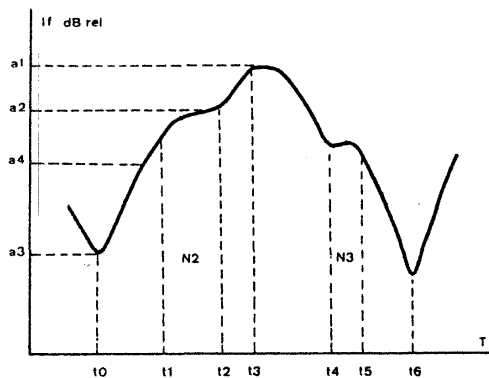
Step one aims at an initial 'broad' segmentation on the basis of dips in the intensity of the band-pass filtered speech. In order to determine the importance of local dips, Mermelstein's convex-hull method (ref 4) is applied to the data in a window of variable length (initially 500ms long). If the intensity difference at the largest dip exceeds the 2 dB threshold, the window is narrowed so that the dip becomes the rightmost point in the window. If no such dip is found, a segment is formed whose boundaries coincide with the window. The nucleus within the segment is defined as the data points which are no more than 3 dB below the local peak. Segments are rejected only when the peak of the nucleus is below the absolute level threshold.

Step two attempts further segmentations of the initial segments in the regions between the nucleus and the broad syllabic boundaries. This step is applied only to onsets or codas of at least 90 ms. The onset region is scanned for a possible CV transition as manifested by a change in the slope of the intensity curve. An extra nucleus is formed between the leftside boundary of the syllable and the transition point.



Automatic segmentation into syllables for the sentence 'exceptionnel me paraît un grand mot, mais disons que j'ai un destin un peu différent ...' spoken by a female speaker. The rms intensity of the 0.5-3 kHz band (If) is displayed in the lower part of the figure, together with temporary segments after step 2. The rms intensity of the 0-4 kHz band (Ig) appears in the central part, together with syllabic nuclei after step 3. Zero-crossing rate is shown as the dashed line in the same window as Ig. Parameter ranges are indicated between brackets on the status line above the figure. Calibration marks on the time axis are 100 ms apart.

Syllables, numbered from 1 to 20, are also shown by ad hoc symbols (with the rectangular box corresponding to the nucleus and vertical lines to syllable boundaries) beneath the phonetic transcription. Step 2 permitted the detection of syllables 5 and 11, not separated from adjacent nucleus by a clear dip (>2dB), but also introduced the illegal syllable 9. Nucleus readjustment was effective for syllables 1, 5, 7, 10, 12, 16, 17, 19 and 20. Temporary segments between 1 and 2, 2 and 3, 7 and 8, 15 and 16, 18 and 19 are eliminated by step 3.



a1 peak of nucleus N0 of step 1
 a2 peak of additional nucleus N2
 a3 value of If at left boundary
 a4 $a4 = a3 + (a1 - a3) / 2$
 t0 left boundary of segment N0
 t1 left boundary of N2
 t2 right boundary of N2
 t3 peak of N0
 t4 dip ($< 2\text{dB}$), left boundary of N3
 t5 right boundary of N3
 t6 right boundary of segment N0

This extra nucleus is accepted only if the following conditions are met:

1. $a2 > a4$
2. $a1 - a2 < 10 \text{ dB}$
3. $a2 > \text{absolute level threshold for } I_f$
4. $t2 - t1 \geq 35 \text{ ms}$

If a dip ($< 2\text{dB}$) is found in the coda region, a third nucleus is formed after the dip. This nucleus is accepted if $t5 - t4 \geq 35\text{ms}$.

Step three takes as its input I_g as well as the output of step two.

Segments may be discarded on several grounds. (1) Fricativelike segments are detected on the basis of the zero-crossing rate. (2) The peak level of the nucleus must exceed the absolute level threshold. (3) The intensity drop between the level of the previous segment and the level of the nucleus must be smaller than 10 dB. (4) Finally, voiceless segments, as detected indirectly from the F0 data, are rejected.

The nucleus of the remaining segments is readjusted in order to correct the initial segmentation. If the intensity drop between the peak inside the nucleus and the right syllable boundary exceeds 6 dB or 15 dB, the right side of the nucleus is set to the point 6 dB resp. 15 dB below the level of the peak. No adjustment is made for small drops, since this would lead to the inclusion of the postvocalic consonant. Adjacent segments may also be merged under certain conditions. The required minimum duration of the nucleus after readjustment is 25 ms.

RESULTS

The algorithm has been tested on continuous speech at various speech rates, with nuclei as short as 25 ms. Most errors are related to oversegmentation. Long stressed syllables frequently manifest spectral changes at pitch changes or at VC transition. These changes in timbre are reflected by dips in the I_f curve. Occasionally omission errors occur between short unstressed vowels separated by [l] or [r].

1. J-Y. Gresser & G. Mercier (1975), in: G. Fant & M.A.A. Tatham (eds.), Auditory analysis and perception of speech. N-Y.:Academic Press, 349-359.
2. W.A. Lea et al. (1975), IEEE Trans. ASSP-23, 30-38.
3. Ph. Martin (1979), in: H. Hollien & P. Hollien (eds.) Current Issues in Linguistic Theory 9(2), 1091-1094.
4. P. Mermelstein (1975), J. Acoust. Soc. Am. 58(4), 880-883.
5. A.C.M. Rietveld (1984) Syllaben, klemtonen..., Ph.D. Univ. Nijmegen.
6. G. Ruske & T. Schotola (1977) Proc. 9th Int. Congr. Acoustics (Madrid, 4/9-VII-1977), vol. 1., 489 (I-83).
7. G. Ruske & T. Schotola (1982), in: J-P. Haton (ed) Automatic speech analysis and recognition, Dordrecht:Reidel, 153-163.
8. D.C. Sargent et al. (1974), J. Acoust. Soc. Am. 55(2), 410 (A).
9. A.N. Stowe (1963), J. Acoust. Soc. Am. 25, 806 (A).
10. C.J. Weinstein et al. (1975), IEEE Trans. ASSP-23, 54-67.
11. E. Zwicker et al. (1979), J. Acoust. Soc. Am. 65(2), 487-498.