AN APPROACH FOR AUTOMATIC DETERMINATION OF BREAK POINTS IN THE SPEECH WAVEFORM

S. Seidl - F. Poirier
CNRS - URA820

Telecom Paris - Departement Signal
46 rue Barrault - 75013 Paris Cedex 13

ABSTRACT

Very often in signal analysis it is necessary to determine certain points in the signal waveform as anchor points for the processing.

This paper proposes a strategy to find automatically the voice onset for initial vowels, the burst point for voiceless and voiced stops and the release point for nasals.

The presented approach exploits only the information inherent in the waveform of the signal by imitating the human skill of abstraction. The procedure is divided into the following steps:

1. data compression by
   a) coding the original signal sampled at 16 kHz by its local extrema and their position in time
   b) eliminating the silent parts of the signal
   c) signal smoothing
2. classification of the syllables by their type
3. special treatment for each type of syllable

1. INTRODUCTION

For many applications in speech processing a kind of preprocessing is performed to derive some a priori information of the signal. For example, in case of speech recognition the signal has to be segmented intelligently what can only be done by using a coarse classification of the sequence of signal units and looking for some logical borders.

This work has been inspired by a study of S. Kitazawa [1] where he applied data analysis techniques on the results of spectral analysis centered on a specific point in the speech signal - voice onset for initial vowels, burst point for voiceless or voiced stops and release point for nasals. These points were determined by visual inspection of the signal waveform.

The test data base consists of 4200 syllables which are a combination of one of the consonants / ?, p, t, k, b, d, g, m, n, f, v followed by one of the vowels / a, o, e, e, u, y, i, ë, ë, ö /. They are pronounced by 40 french male speakers.

As the results of this analysis are very promising it is now necessary to find automatically the anchor points.

The strategy which is presented in this paper analyses and imitates the human ability of abstraction and universal perception. For that reason only the signal waveform is used to determine the searched points.

At first the signal is smoothed to suppress "noisy variations" of the signal amplitude and to reveal the "main variations", a kind of signal periodicity and the envelope of the waveform. Then the syllables have to be classified by their type - isolated vowel, initial voiceless stop, initial voiced stop or initial nasal - because the characteristics of the anchor point depend on the type of the consonant. That is why a special search strategy has been developed for each type of syllable.

1. In case of an isolated vowel the problem is reduced to finding the beginning of the syllable.
2. In case of an initial voiceless stop the beginning of the vowel is indicated by growing signal amplitude and increase in periodicity.
3. In case of an initial voiced stop the signal corresponding to the consonant is very regular. The beginning of the vowel is characterized by a break point in the signal waveform and by growing amplitude.
4. In case of an initial nasal the anchor point can only be determined by regarding the behavior of the signal envelope. As the transition between a nasal and a vowel is rather smooth the position of the resulting point is a priori very uncertain.

2. SUMMARY OF THE WHOLE PROCEDURE

This section will give a brief summary of the sequence of processing steps for the whole search procedure.
1. The original signal is sampled at 16 kHz and coded by 16 bits.
2. In order to associate the representation of the signal with the visual impression it is coded by its local extrema and their position in time.
3. Now the silent parts between the syllables can be eliminated without losing the information of their position in time.
4. As not all the extrema influence equally the visual impression of the signal of a syllable a method has been developed to extract only the "most important" ones. This treatment will be called signal smoothing. It results in coding the signal by its "main" extrema and their position in time (see section 3.).
5. In order to determine the type of the syllable a characteristic vector representing only information of the smoothed signal is calculated for each syllable (see section 4.).
6. The classification is based only on the characteristic vectors and uses a cluster algorithm (see section 4.).
7. Depending on the results of the classification a special treatment has to be applied for each type of syllables (see section 5.).

3. SIGNAL SMOOTHING

Signal smoothing in the context of the presented procedure means extraction of "main" extrema. As justification for the selection of certain extrema the smoothed signal can be visualized by linear interpolation of the remaining extrema and may be compared with the original signal to verify that only the main tendencies of the signal have been preserved and the noisy break points have been suppressed (see Fig. 2).

The smoothing strategy is explained for the case of a maximum because there is no fundamental difference in the procedures for a maximum or a minimum.

For the decision whether an extremum $e_i$ has to be kept or to be eliminated, a threshold $s$ for the difference between two consecutive extrema was chosen and a window of 5 amplitudes $[e_{i-2}, e_{i-1}, e_i, e_{i+1}, e_{i+2}]$ was examined. After having calculated

$$dl = e_{i-1} - e_i$$
$$dr = e_i - e_{i+1}$$

there are four cases to be distinguished. Only in the case of $dl > s$ and $dr > s$, $e_i$ is supposed to be a main maximum (see Fig. 1a). In any other case the tendency of the signal in context has to be examined.

Therefore a kind of slope is calculated

$$pl = \text{sig}(e_{i-2} - e_i)$$
$$pr = \text{sig}(e_i - e_{i+2})$$

In Fig. 1 the extrema which will be eliminated are marked with a circle.

Fig. 1: Different cases of signal smoothing
After having treated a frame of amplitudes its position is shifted by two amplitudes only, which does not mean that the procedure is redundant because the meaning of each extremum for the selection process depends on its position in the examined window.

Fig. 2 shows an example for a so smoothed signal in comparison with the original signal.

![Fig. 2: Signal of "BA", 62 ms around the break point, before and after smoothing](image)

4. CLASSIFICATION

The classification of the syllables by their type is done hierarchically in two steps. The first partition separates the isolated vowels and the syllables with initial voiceless stops from the syllables with initial voiced stops or nasals. In the second step the two resulting groups are subdivided independently into final classes. Each decision step is based on the detection of special signal properties. So the first distinction may be made by examining the very beginning of the syllables. The further decisions have to take into account the evolution of the signal of the syllable.

The 15 parameters of the characteristic vectors are the results of the analysis of two parts of the smoothed signal. The first one is a frame of 12 ms at the very beginning of the syllables; the second one is a frame of 62 ms after the beginning of the syllables. For both of the frames the number of resting extrema is taken as information of regularity and periodicity. The mean and variance of the maxima and minima give an idea of signal energy. To estimate the regularity of the signal the mean of the differences between consecutive maxima is calculated as well as for consecutive minima. The last of the 15 parameters is the duration of the syllable.

To reduce the number of parameters the most significant ones have to be detected for each step of the classification by discriminant analysis. The classification itself is done by means of clustering. Here an algorithm of K nearest neighbours is used.

5. SPECIAL TREATMENT FOR EACH TYPE OF SYLLABLES

1. Isolated vowels
   In this case the anchor point coincides with the beginning of the syllable and the problem is reduced to detecting the silent parts of the signal. This is done by using a threshold of 300 for the amplitude and supposing a minimum duration of 0.25 s for each syllable.

2. Initial voiceless stops
   Like in the preceding case the anchor point is the beginning of the syllable. Besides, the beginning of the vowel could be detected by looking for a sudden increase of the signal amplitude. To make sure having found the vowel the further evolution of the signal is examined.

3. Initial voiced stops
   The signal of a voiced stop is very regular and periodic so that after signal smoothing the number of resting extrema in a certain time interval of 6 ms is much less for the consonant than for the vowel. The break point indicating the beginning of the vowels is identified by a sudden increase of the number of extrema which shouldn't be less than 100 ms, the supposed minimum duration of a voiced stop, after the beginning of the syllable.

4. Initial nasals
   The treatment of syllables with initial nasals is the most difficult one. Only the envelope and the zero crossing rate give indications of the anchor point. To find a representation of the envelope the signal was cut into intervals of 12 ms. Then the biggest as well as the smallest amplitude of each interval were selected to represent the envelope. Besides the zero crossing rate for each interval was calculated. The anchor point has been defined as the point where a significant increase in the zero crossing rate or in the envelope amplitude is detected before the
maximum amplitude of the syllable which is supposed to belong to the vowel.

6. EXPERIMENTS

1) Classification of the syllables

For separating the isolated vowels and the syllables with initial voiceless stops from the syllables with initial voiced stops or nasals the most significant parameters are

a) the number of extrema in the first window
b) the duration of the syllable
c) the mean of the differences between consecutive maxima of the second window.

The partition based on these parameters reaches a score of 96.2%.

For the identification of the type of syllables with either initial voiced stops or nasals only the variance of the maxima of the second window is examined. Finally the type for 82.7% of all the syllables could be determined correctly. These results support the hypothesis leading to the choice of parameters for the characteristic vector.

2) Determination of the anchor point

To give an idea of the quality of the presented strategy the automatically obtained points $x_A$ are compared with the anchor points $x_V$ determined by visual inspection of the signal. The following histograms show the distributions of the variable $e = x_V - x_A$ for each type of syllables separately. The type was supposed to be known a priori.

Fig. 3: Distribution of $e$

7. CONCLUSION

The presented method to determine break points in the speech signal exploits only waveform information. It tries to imitate the human skill of abstraction. Therefore at first a kind of signal smoothing is performed to reduce the variability of the waveform and then after a classification of the syllables a special search procedure is applied depending on the type of the syllable.

Only in the case of syllables with initial nasals the automatically determined anchor points differ seriously from the visually determined ones. But as the transition between nasals and vowels is highly coarticulated the existence and position of a break point is a priori uncertain.

8. REFERENCES

