# SYLLABLE DETECTION IN READ AND SPONTANEOUS SPEECH

*Hartmut R. Pfitzinger, Susanne Burger, Sebastian Heid*

hpt|burger|heid@phonetik.uni-muenchen.de
Institut für Phonetik und Sprachliche Kommunikation
University of Munich, Germany

## ABSTRACT

Automatic syllable detection is an important task when analysing very large speech corpora in order to answer questions concerning prosody, rhythm, speech rate, speech recognition and synthesis. In this paper a new method for automatic detection of syllable nuclei is presented. Two large spoken language corpora (PhonDatII, Verbmobil) were labelled by three phoneticians and then used to adjust the key parameters of the algorithm and to evaluate its error rate. Additionally, parts of the corpora were used to test the inter- and intra-individual consistency of the transcribers. The evaluation of the algorithm currently shows an error rate of 12.87% for read speech and 21.03% for spontaneous speech. The inter-individual consistency of 95.8% might be considered as an upper limit for any automatic detection method.

## 1 INTRODUCTION

Undoubtedly the syllable has proven to be an important concept in theoretical description of spoken language. But there is still neither a clear theoretical definition nor a perfect way for a practical determination of actually spoken syllables. As Tillmann 1964 [4] has shown there are many independent and even contradictory theoretical approaches. Pre-theoretic intuitions of speech scientists have only inspired but not determined the theoretical approaches which have failed to lead to a definition that is useful in practice and empirically testable.

We avoided an important source for discussion by omitting the question of syllable boundaries, concentrating instead on finding the syllable nuclei. As a point of departure we considered the following observations: (i) all speech scientists seem to have a clear intuition of what a syllable is, and (ii) loudness which is a direct result of mouth opening and also a primary cue of sound perception, seems to be a good feature for a signal based syllable detector.

The next steps were, (i) to test the intuition of phoneticians by letting them label syllables, and then to test the inter- and intraindividual consistency, and (ii) to develop a loudness based automatic syllable detector and to evaluate its output referring to the manually labelled speech data.

### 1.1 Speech material

The read speech material was taken from the PhonDatII speech database. PhonDatII contains 200 sentences dealing with railway information queries read aloud by 16 speakers (10 male, 6 female), recorded in three German regions (Kiel, Bonn, München). For each of the 16 speakers 60 sentences have been chosen resulting in 70 minutes of speech.

A subset of the Verbmobil database [5] consisting of negotiation dialogues served as the spontaneous speech corpus. We used the CD-ROM 7.0 containing 68 dialogues (1739 turns) recorded in the regions mentioned above. We used 52 turns, 16 from Kiel spoken by 2 male speakers, 18 from Bonn spoken by 5 female and 3 male speakers, and 18 from München spoken by 4 female and 4 male speakers, resulting in 3177 manually labelled syllables (15 minutes).

### 1.2 Manual labelling

The 960 sentences of read speech have been segmented manually by three phoneticians, which resulted in 14048 syllables. In addition, each of the phoneticians re-labelled 96 sentences four months after the first session. To promote greatest consistency only one phonetician labelled the spontaneous speech corpus. The segmentation convention was to mark the nuclei of perceptible syllables near the peak of the loudness envelope. In undecidable cases the segmenters had the possibility to label with a question mark. This was done for 1% of all syllable marks.

## 2 AUTOMATIC DETECTION METHOD

In a syllable based automatic speech recognition system Weigel [6] relied on a modified Bark-based loudness estimation for syllable nucleus detection. His experiments showed that the energy peaks in the frequency range from 250 Hz to

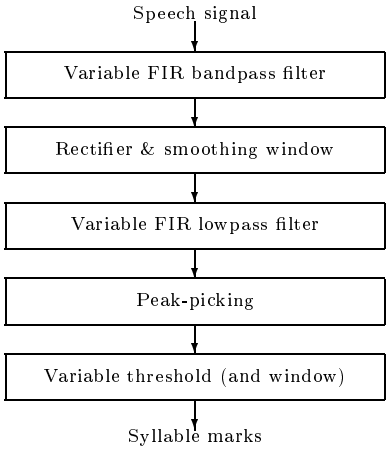


Figure 1: Scheme of the algorithm for syllable detection.

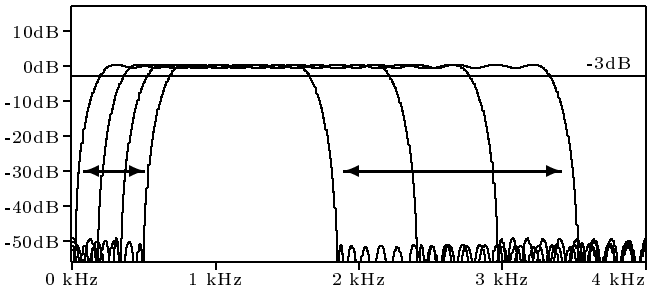


Figure 2: The possible cutoff frequencies of the bandpass filter are 200, 360, 520, and 680 Hz at the low band-edge and 1650, 2210, 2770, and 3330 Hz at the high band-edge.

2500 Hz are well-correlated with the syllable nuclei. Consequently our method (see Fig. 1) estimates the logarithmic short-term amplitude of the bandpass filtered speech signal. Fig. 2 shows the possible transfer functions of the bandpass filter. Preliminary experiments revealed the range of cutoff frequencies to be covered. Lowpass filtering of the logarithmic short-term amplitude (Fig. 4) was applied to suppress ripples caused by F0 or transient phonemes and to force the system to oscillate at the syllable frequencies. Therefore the cutoff frequency range of 7 to 13 Hz was found by analysing syllable distances (see Fig. 12). The peaks of the resulting energy contour served as candidates for syllables.

### 2.1 Threshold post-processing

To decide whether a candidate is a syllable or not we first used a simple threshold criterion. As illustrated in Fig. 3 we divided all candidates into two groups: in the upper histogram we counted the candidates that matched the reference syllables as a function on the signal loudness. In the lower

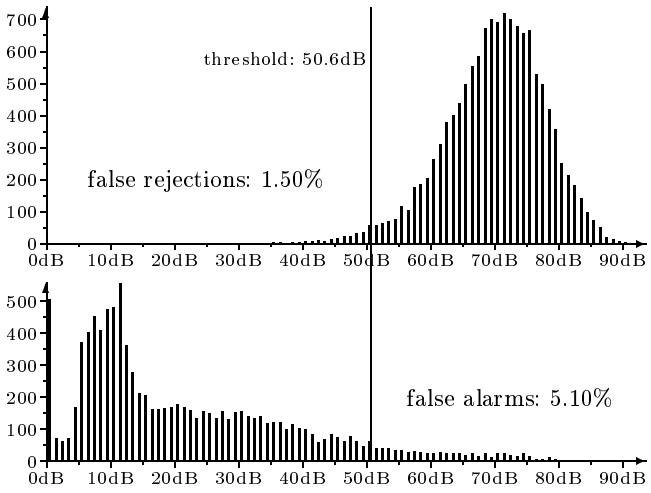


Figure 3: Energy distribution of the candidates that match the reference syllables (upper) and that do not (lower). The error rate of 12.87% is the sum of false rejections, false alarms, and misses of reference syllables (6.27%).

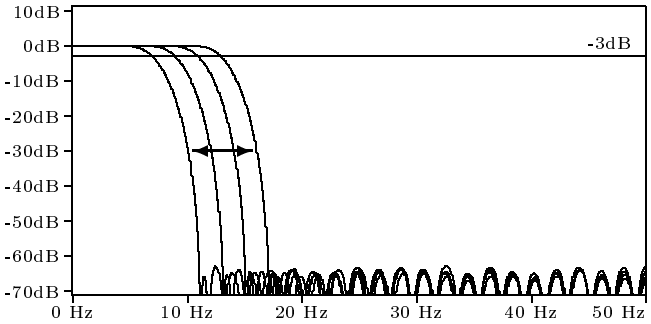


Figure 4: The possible cutoff frequencies of the lowpass filter are 7, 9, 11, and 13 Hz.

histogram the candidates are counted that re-mark a just marked syllable again or that are farther away than 100 ms from any reference syllable. The candidates on the left side of the threshold line were rejected and the ones on the right side were accepted. This yields two classes of errors: false rejections and false alarms. A third class of error is the number of reference syllables that were missed by any candidate.

### 2.2 Window post-processing

The large percentage of false alarms is produced by multiple marking of identical syllables. To reduce this source of errors we placed a window over every candidate and removed all candidates that fell inside the window except the candidate with the highest loudness. Fig. 5 illustrates that a window size of 88 ms can decrease the total error rate by 1.6% for read speech and by 2.7% for spontaneous speech (Fig. 6). The increase of the error rate at larger windows is caused by the increasing elimination of short syllables.

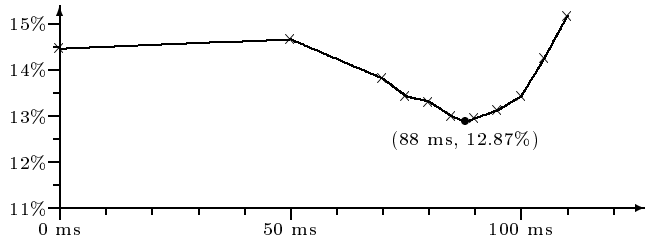


Figure 5: Error rate for read speech as a function of the window size.

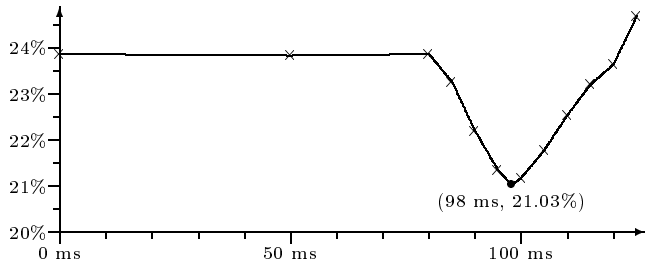


Figure 6: Error rate for spontaneous speech as a function of the window size.

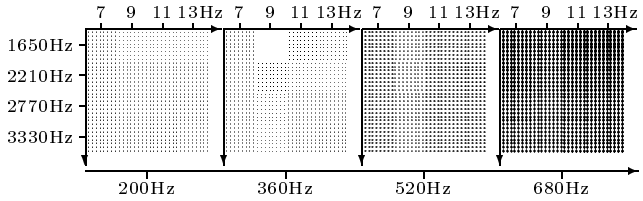


Figure 7: Error rates of read speech syllable detection as a function of bandpass low cutoff (*bottom x-axis*), high cutoff (*y-axis*), and lowpass cutoff (*top x-axis*) (white=12.87%, dark=24.12%).

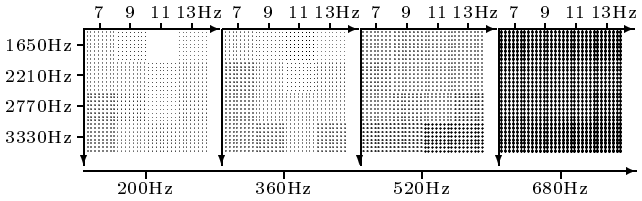


Figure 9: Error rates of spontaneous speech syllable detection as a function of bandpass low cutoff (*bottom x-axis*), high cutoff (*y-axis*), and lowpass cutoff (*top x-axis*) (white=21.03%, dark=31.19%).

## 3 RESULTS

Five parameters of the detection algorithm are variable: three cutoff frequencies, the threshold, and the window size. A training procedure minimizes the error rate for each combination of the cutoff frequencies by optimizing analytically threshold and window size. The resulting error rates for the read speech corpus are depicted in Fig. 7. A minimum error rate of 12.87% (white square) resulted from a bandpass filter from 360 to 1650 Hz and a lowpass filter with a cutoff frequency of 9 Hz. For spontaneous speech a bandpass filter from 200 Hz to 1650 Hz and a 11 Hz lowpass filter led to the minimum error rate of 21.03% (Fig. 9). In order to determine the sufficient training corpus size for maximum variability and enough redundancy to enable generalization, the number of training sentences for each evaluation was increased from 60 to 960. Fig. 8 shows that a corpus size of about 500 sentences leads to the maximum error rate. A further increase of the number of sentences only adds redundancy, so that the error rate goes slightly down.

### 3.1 Consistency of manual labelling

The three segmenters re-labelled 96 sentences of the read speech corpus four months after the first session to evaluate the inter- and intraindividual consistency. If marks are within a 50 ms window they were considered to mark the same syllable. In Tab. 1 the intraindividual consistency is shown. E.g. Segmenter no. 2 matched 98.5% of his former syllable marks with a very small mean standard deviation of 4.14 ms. Tab. 2 summarizes the interindividual consistency of the three segmenters. They agreed in 95.8% of all marks. Two segmenters agreed in 2.1% of all marks, and another 2.1% was set by only one segmenter.

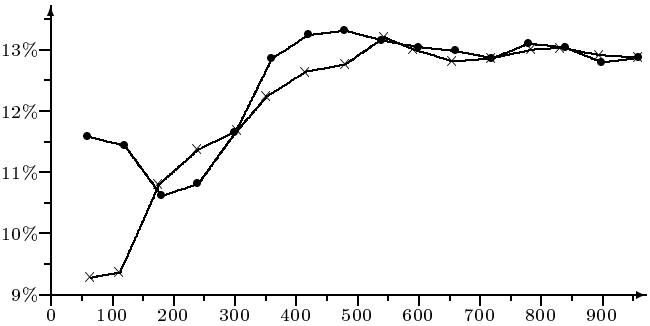


Figure 8: Error rate as a function of the number of training sentences. ×: 16 speakers, increasing number of sentences. •: 60 sentences, 1 to 16 speakers.

### 3.2 Problems

Generally, two types of labelling problems can be distinguished: (i) the phoneticians agreed in labelling a syllable which the system didn't find and so it is a system fault, (ii) the phoneticians disagreed in labelling a syllable. Problems of type (i) are rather simple to handle but reveal insufficiencies of the system. The main shortcoming showing up is the incapability of the system to handle strong variations of speech rate. So methods using adaption to different speech rates should be a promising field for future research. Problems of type (ii) are difficult to deal with because in many cases there is no way to decide whether there is a syllable or not.

Problems of type (ii) can be subdivided into two categories. Firstly, even very clear intuitions of the segmenters do not seem to correspond to the signal. In such cases often higher-level knowledge, like the phonotactics of his own language, shaped the segmenter's intuition. Secondly, the segmenters could not make a clear decision because they met an ambiguous acoustic structure.

A typical example for the first category are collisions of

| segmenter | self match (within 50 ms) | mean std. dev. |
|---|---|---|
| no. 1 | 96.1% | 5.58 ms |
| no. 2 | 98.5% | 4.14 ms |
| no. 3 | 96.7% | 6.44 ms |

Tab. 1: Intraindividual consistency of each of the three segmenters.

| agreement | syllables | percent | mean std.-dev. |
|---|---|---|---|
| 3 segmenters | 2461 | 95.8% | 5.79 ms |
| 2 segmenters | 54 | 2.1% | 11.63 ms |
| 1 segmenter | 55 | 2.1% | — |

Tab. 2: Interindividual consistency of three segmenters each labelling the same 96 sentences of the read speech corpus.

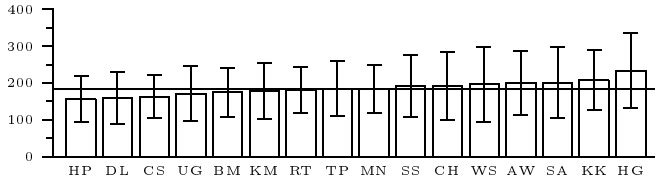


Figure 10: Average syllable distances of the speakers from the read speech corpus. Average over all speakers $\bar{x} = 184.1$ ms, standard deviation $\sigma = 81.1$.

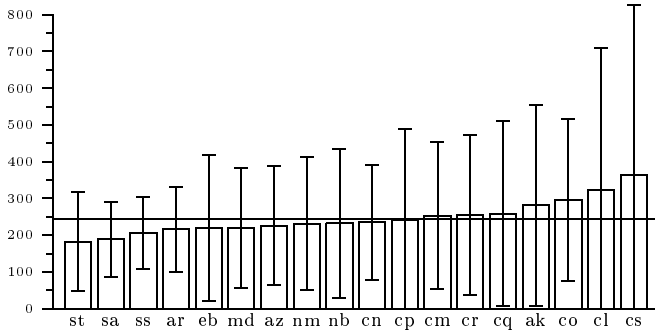


Figure 11: Average syllable distances of the speakers from the spontaneous speech corpus. Average over all speakers $\bar{x} = 244.9$ ms, standard deviation $\sigma = 242.4$.

vowels at syllable boundaries as in "...ohn*e* *U*msteigen". The /e/ and the /u/ form a diphthong which is clearly perceptible as two syllables because a one-syllable-segmentation is incompatible with the German phonotactics, but the peaks of these two syllables seem to meet in one point.

A typical example for the second category are reduced syllables which are assimilated to one syllable. Between a clear two-syllable-case and a clear one-syllable-case all grades of assimilation are possible. For these grades there is no unambiguous decision criterion.

### 3.3 Speech rates

To get a first impression about the applicability of our method we measured the temporal relations of the syllable nuclei. We compared results for both read and spontaneous speech. At first sight it is surprising that the average distances of read speech nuclei are shorter than those of spontaneous speech and that the spontaneous speech nuclei distances have a considerable higher standard deviation than the read speech nuclei distances (see Fig. 10 and 11).

In Fig. 12 we collected a) all read speech nuclei distances and b) the spontaneous speech nuclei distances shorter than 600 ms from all speakers. The histograms look quite similar but there are a lot more distances over 300 ms in the spontaneous speech material. The fact that (i) the average distance is roughly 60 ms longer, that (ii) the standard deviation is higher, and that (iii) inter-nuclei distances of more than 300 ms are more frequent in the spontaneous speech material has to be seen in the light of typical phenomena of the given dialogue situation: long pauses as well as hesitations in the form of filled pauses and segment lengthenings. Also emphatical accents considerably spread the distances between the nuclei. We found that in spontaneous speech pauses significantly distort the mean distance of syllable nuclei.

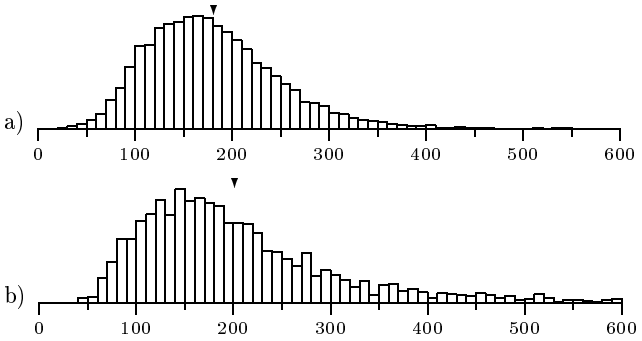


Figure 12: Distributions of the syllable distances in a) the read speech corpus and b) the spontaneous speech corpus. The triangular marks indicate the mean values (only taking into account distances below 600 ms).

## 4 DISCUSSION

In this paper a new approach to syllable detection was presented. We discussed the results and problems of automatic and manual syllable detection and showed the inter- and intraindividual consistency in labelling read speech.

It will be an interesting part of our future research to check the consistency of hand-labelled data of spontaneous speech. Generally, it is found that there can be no 100% detection, because there is no 100% criterion which might define a syllable. This allows the conclusion that any linguistic syllable characterization should, in actual application, be abandoned in favour of an operationally defined acoustic quasi-syllable.

## REFERENCES

[1] H. J. Cedergren and H. Perreault. Speech rate and syllable timing in spontaneous speech. In *Proceedings of the ICSLP 94*, volume 3, pages 1087–1090, Yokohama, 1994.

[2] T. H. Crystal and A. S. House. Articulation rate and the duration of syllables and stress groups in connected speech. *JASA*, 88:101–112, 1990.

[3] W. Reichl and G. Ruske. Syllable segmentation of continuous speech with artificial neural networks. In *Proceedings of Eurospeech '93*, volume 3, pages 1771–1774, Berlin, 1993.

[4] H. G. Tillmann. *Das phonetische Silbenproblem — Eine theoretische Untersuchung.* PhD thesis, Univ. Bonn, 1964.

[5] H. G. Tillmann et al. The phonetic goals of the new bavarian archive for speech signals. In *Proceedings of the XIIIth ICPhS*, volume 4, pages 550–553, Stockholm, 1995.

[6] W. Weigel. *Silbenorientierte Erkennung fließender Sprache mittels diskreter stochastischer Modellierung.* PhD thesis, TU München, 1990.