There are unquestionably a number of improvements which still need to be made to the model but these may not become obvious until it has been applied to sufficient examples of elliptical speech. It works for the instances tried above but it remains to be seen how wide an application it has. The model was described at the beginning of this paper as a preliminary one – this in anticipation of the many adjustments which will no doubt be made to it in later stages.

REFERENCES

Carberry, S. 1989. 'A pragmatics-based approach to ellipsis resolution'. *Computational Linguistics* 15, 75-96.

Firth, J.R. 1957. *Papers in linguistics 1934-51*. London: Oxford University Press.

Halliday, M.A.K. 1961. 'Categories of the theory of grammar'. *Word* 17, 241-292.

Halliday, M.A.K. 1985. *An introduction to functional grammar*. London: Edward Arnold.

Joos, M. 1962. 'The five clocks'. *International Journal of American Linguistics* 28, 9-62.

Lobeck, A.C. 1987. *Syntactic constraints on VP ellipsis*. Bloomington: Indiana University Linguistics Club.

Lyons, J. 1968. *Introduction to theoretical linguistics*. Cambridge: Cambridge University Press.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Quirk, R. & J. Svartvik, eds. 1980. *A corpus of English conversation*. Lund: Gleerup.

Lund University, Dept. of Linguistics
Working Papers 35 (1989), 53-61

Algorithms for Automatic Segmentation of Speech

Lars Eriksson

INTRODUCTION

A problem in speech recognition and also in automatic phonetic transcription from read speech is accurate segmentation of the incoming speech signal into syllable-sized segments. Several methods for automatic speech segmentation using computers have been developed. One common and also quite simple algorithm is to use the intensity from the original signal and the intensity from one or more bandpass filtered versions of the signal. These are compared using different criteria to determine the syllabic boundaries in the speech signal.

The syllable is linguistically defined as "a sequence of speech sounds having a maximum or peak of inherent sonority (that is apart from factors such as stress and voice pitch) between two minima of sonority" according to Robins 1966. The physical parameter that measures sonority would be intensity, and intensity could thus be used for segmentation of speech into syllables. The segments produced by a segmentation based on intensity are, however, not always the same as the syllables defined on a phonological level. The intensity segmentation often locates boundaries that cannot be identified as phone boundaries or even word boundaries.

Mermelstein shows in his work on the convex hull speech segmentation method that it is more convenient to talk about 'syllabic units' rather than syllables when working with automatic segmentation. A syllabic unit consists of the syllable nucleus, that is, the vowel, but may also contain consonants from other neighbouring syllables, especially from intervocalic consonant clusters. For example the words and syllables, *piles of* [pajlz ɔf] would form the syllabic units [pajl] - [zɔf].

One method for speech segmentation using unfiltered speech intensity and bandpass filtered speech intensity has been developed at Lund. This method is used within the project 'Prosodic Parsing for Swedish Speech Recognition', sponsored by the National Swedish Board for Technical Development (STU). The method, which hereafter will be referred to as the filtered speech method, has proved to be reasonably accurate for the purpose of finding syllabic units in normal reading-rate speech. Each syllabic unit contains one 'tonal segment', that is, a vowel and the following sonorants, if there are any. Those tonal segments are used in the project mentioned above for recognition of stressed and unstressed syllables and Swedish word accents. The syllabic unit is therefore in this application a more natural and useful unit than the syllable, because a tonal segment does not always coincide with a syllable but with a syllabic unit. The use of tonal segments in the word accent recognition programme has been described by House et al. 1987a, 1987b, 1988.

ALGORITHMS

The segmentation algorithm used in the word accent recognition project is similar to methods decribed by Mertens 1987. Other algorithms have been described by Mermelstein 1975, Lea 1980, and Blomberg & Elenius 1985.

Mertens proposes a technique which uses the syllable nucleus as a starting point for the segmentation. This is detected as the place where the intensity function has a maximum. A maximum in the intensity function is taken as a syllable nucleus if it is higher than 40 % of the average intensity of the speech signal. The syllable is then formed around this maximum and is defined as the segment where the intensity exceeds a certain threshold value.

FILTERED SPEECH METHOD

Mertens' method has been modified in our implementation, which works like this:

The speech signal is first low-pass filtered at 4 kHz to prevent aliasing distortion and is then sampled at 10 kHz. An intensity function is obtained from this signal using the root-mean-square formula. This is done with the ILS signal processing programme package on a DEC VAX 11/730 mini-computer. This intensity function is referred to as the *unfiltered* intensity function. An additional intensity function is obtained from a digital bandpass filtered version of the sampled signal, from 500 Hz up to 4 kHz, using a filter with 72 dB/octave attenuation. This function is referred to as the *filtered* intensity function. Both intensity functions are smoothed with a moving average filter.

The bandpass filtering is done to reduce the effect from voiced consonants, as they show a concentration of energy in the spectral region below 500 Hz.

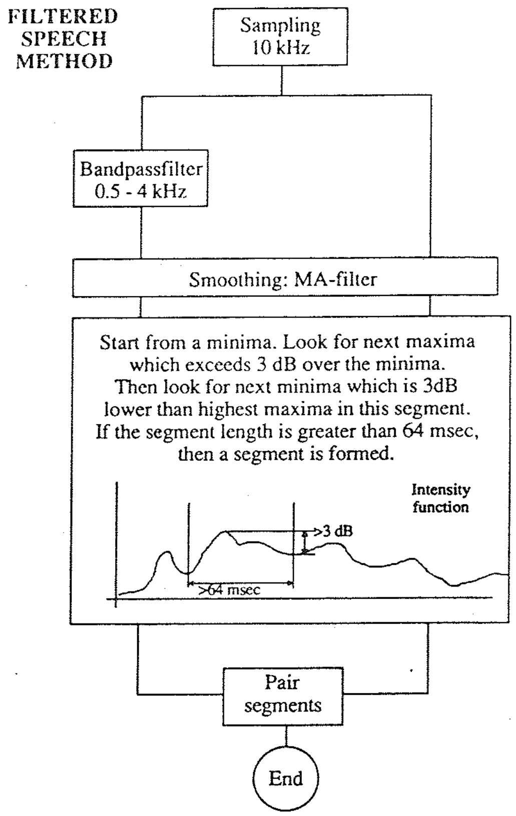

FILTERED SPEECH METHOD
Sampling 10 kHz
Bandpassfilter 0.5 - 4 kHz
Smoothing: MA-filter
Start from a minima. Look for next maxima which exceeds 3 dB over the minima. Then look for next minima which is 3dB lower than highest maxima in this segment. If the segment length is greater than 64 msec, then a segment is formed.
Intensity function
>3 dB
>64 msec
Pair segments
End

Figure 1. The Filtered Speech Method, flowchart.

If this band is attenuated in the speech signal, the voiced consonants will show up as minima in the intensity function.

The next step in the segmentation procedure is a syllabic segmentation algorithm which is applied to both intensity functions, where local intensity minima are used as segment boundaries.

Taking the first minimum as the first boundary, the programme searches for the next maximum which exceeds 3 dB over the intensity level of the preceding boundary. From this maximum the next minimum which meets the following two conditions is taken as the next segment boundary: 1) The intensity difference between the minimum and the highest preceding maximum in the segment must be larger than 3 dB, and 2) The duration from the previous boundary to the minimum in question must be greater than 64 ms. This routine is applied to both intensity functions. The two functions are then compared and the segment boundaries which are closer together than 64 ms are collapsed into one boundary which is placed halfway between the two original boundaries.

The value 3 dB is a rule-of-thumb value we have chosen from our preliminary experiments. This value, and also the 64 ms minimum segment length, may be altered to tune the algorithm.

CONVEX HULL METHOD

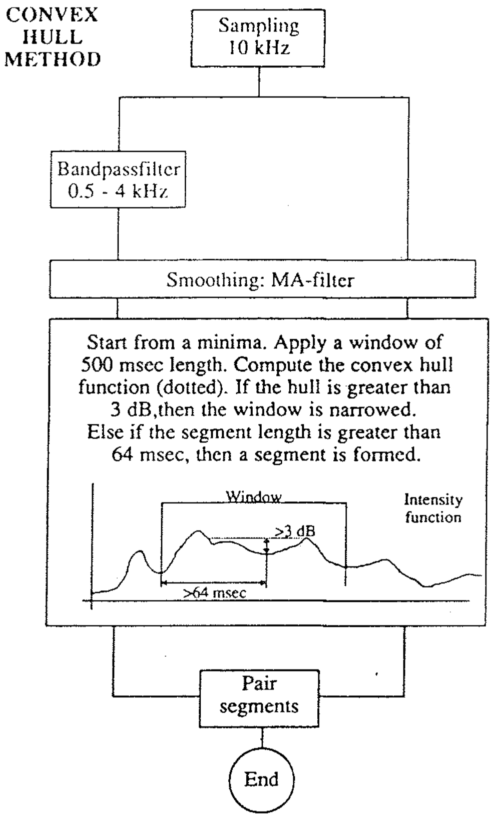
The convex hull method, described by Mermelstein 1975, has also been implemented in software and used in the word accent recognition programme for speech segmentation into syllabic units and tonal segments. The convex hull method may be explained as follows:

First, the first minima in the intensity function is detected. This is taken as the first segment boundary. Then a window of 500 ms is applied to the intensity function, starting from the segment boundary. A new function, the 'convex hull', is computed in this window, which is monotonically nondecreasing from the start of the window to the point of maximum intensity, and monotonically nonincreasing from the maximum to the end of the window.

Then, the maximum distance between the intensity function and the convex hull function in this window is searched for. If this distance exceeds a certain threshold value, for example 3 dB, then the window is narrowed so that the end of the window is placed at this point, and the convex hull function is then computed for this new, shorter window.

If the maximum distance between the convex hull and the intensity function is less than the threshold value, then the end of the window is marked as a segment boundary and the window is moved to the right to begin from this point. The window length is also readjusted to 500 ms.

The convex hull algorithm has been slightly modified in our implementation in the ILS-system. An extra module has been added which adjusts the length of the window before the first pass. This length is originally 500 ms as decribed above, but is adjusted so that the end of the window will coincide with a minimum in the intensity function before computing the first convex hull function. This is done to reduce the effect of false segment boundaries



Start from a minima. Apply a window of 500 msec length. Compute the convex hull function (dotted). If the hull is greater than 3 dB,then the window is narrowed. Else if the segment length is greater than 64 msec, then a segment is formed.

Figure 2. The Convex Hull Method, flowchart.

due to long syllables, for example in final position. Otherwise a segment boundary may be put where actually no segment boundary is to be found. The implementation of the convex hull method in our programme also uses the two-pass procedure decribed earlier, with one *unfiltered* intensity function and one *filtered* intensity function.

EXPERIMENT

Ten sentences of normal reading-rate speech, with two different male speakers, were recorded and analyzed with our two speech segmentation algorithms: our own filtered speech method and the convex hull method. Both methods were tested with different threshold values. Mermelstein proposes a threshold value of 2 dB and we have run our method with a 3 dB threshold, so it would be interesting to test both methods at 2 dB and 3 dB.

The ten test sentences read by speaker number 1 contained 120 syllables in all. One would thus expect to find 120 syllabic units in this material. A syllabic unit was taken as correctly identified if the syllable nucleus fell into the segment that the speech segmentation algorithms produced. The segment was then identified as a syllabic unit.

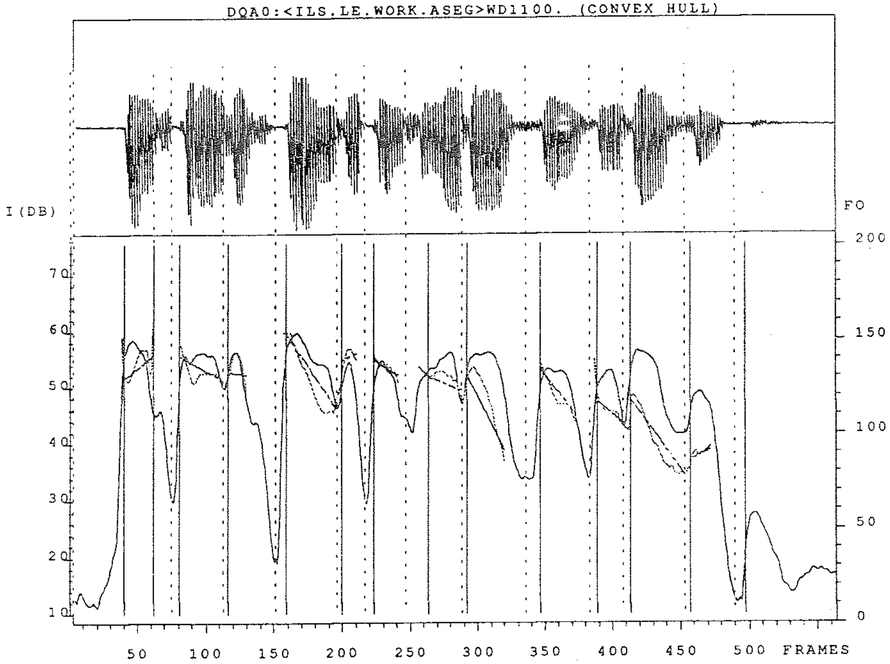


Figure 3. Output from the segmentation programme, using the convex hull method. The sentence segmented is *Hon skriver ständigt om sina resor i sin dagbok* 'She always writes of her travels in her diary'. A segment boundary is marked in the plot with a vertical dotted line. Assumed vowel onset (not discussed here) is marked with a solid vertical line. The lower part of the plot shows: original pitch contour (dotted line), linearized pitch (dashed line) and intensity (solid line). 1 frame is 6.4 ms.

RESULTS

First with a 2 dB threshold, the convex hull method identified 111 of the 120 syllabic units in the speech material. This corresponds to 92%. The filtered speech method identified 113, which is 94% of the syllabic units. With a 3 dB threshold these scores fell to 90% for the convex hull method and 92% for the filtered speech method. The sentences read by speaker 2 contained 118 syllabic units, that is, two less than speaker 1. This was because speaker number 2 made reductions of two syllables. With a 2 dB threshold, both methods identified 105 syllabic units which corresponds to 89%. Increasing the threshold value to 3 dB lowered this identification score to 103 units or 87%. This shows that both methods were equally good, or bad, if one only looks at the number of detected syllabic units.

However, both methods produced extra segments, both voiced and unvoiced. About ten extra unvoiced segments were found in the ten test sentences. Segments with no voicing or less than 20 ms voicing were defined as unvoiced segments. These were detected by both algorithms and at both threshold values but could easily be sorted out from the correctly detected segments. It was mainly the fricatives and plosives, especially [s], and [p], [t] and [k] in final position, that produced these segments. It would be possible to add these unvoiced segments to adjacent voiced segments, thus forming syllabic units. Rules for this should be possible to write, but this has not been considered here, because it falls outside the aim of the Prosodic Parsing Project, where only the voiced part of a segment, the tonal segment, is of interest.

A number of false voiced segments were also detected by the segmentation methods. Looking first at the sentences read by speaker number 1, the convex hull method wrongly detected 6 voiced segments and the filtered speech method 10 voiced segments at the 2 dB threshold. With the 3 dB threshold these values decreased to 2 and 6 respectively. The false segments were produced by sonorants, mainly [n] and [m]. The filtered speech method seemed to produce more false segments than the convex hull method. It also is clear that bandpass filtering from 500 Hz to 4 kHz does not completely eliminate the interference from all sonorants. For speaker number 2, the number of false voiced segments was only 1 for both methods at both threshold values. This is interesting and it seems that speaker 2's more casual way of reading reduced the effects of sonorants in the segmentation. The sentences read by speaker 1, who read more clearly, produced significantly more false voiced segments. The false segments lowered the total identification score to about 88% for both speakers.

The undetected syllabic units were mainly segments with two unstressed vowels and a voiced consonant between, for example *ejö* [ɛjø], *eri* [eri], *are* [arə] and so on. No segment boundary was then put at the voiced consonant. The segmentation also missed the boundary between vowels where one unstressed and one stressed vowel were reduced to one unit, for example *io* [iu] in *i ordet*, *oe* [uɛ] in *poesi* and *ey* [ey] in *e yrkesmålare*.

It is obvious that these methods show sensitivity both to speaker voice quality and to different productions of the same words. The latter problem may be solved in a speech recognition programme with word reduction rules at a phonological level and perhaps also with different production templates for certain words.

An additional test of the algorithms was also performed. The average speed of the segmentation of a sentence was measured. This showed that the filtered speech algorithm was much faster than the convex hull method. The average time per sentence was 94 ms for the filtered speech method and 520 ms for the convex hull method. The programming language that we have used is Fortran. No attempt has been made to tune the code in our programme. Both methods may well run faster, but these first results show that the filtered speech method is the fastest, and this may be an advantage if the programme is to run in real-time.

Table 1. Results from segmentation with 3 dB threshold of the test sentences spoken by two male speakers. (FS = Filtered Speech method, CH = Convex Hull method)

	Speaker 1		Speaker 2	
Total number of segments	120		118	
Method used	FS	CH	FS	CH
Segments identified as				
... correct	110	108	103	103
... false voiced	6	2	1	2
... false unvoiced	10	9	9	7

CONCLUSIONS

These algorithms for segmentation of reading-rate speech work quite well if one takes into account that they are very simple and easy to implement on a computer. The overall accuracy may not be sufficient to make these programmes good final segmentation algorithms for an automatic speech recognition programme. However, they may be used as tools for a preliminary analysis of the segments in the speech, and the segments can then be used for testing rules pertaining accent patterns, performing an automatic stylization of the Fo contour, etc. These are areas in which we have been using these segmentation programmes. We have then gone a few steps further and added rules to find the vowel in a syllabic unit, and we have also extracted pitch and duration information from the syllabic units. This information has been used by the stress and word accent rules in the Prosodic Parsing Project, House et al. 1989.

REFERENCES

Blomberg, M. & K. Elenius. 1985. 'Automatic time alignment of speech with a phonetic transcription'. *Proceedings of the French Swedish seminar on speech*, eds. B. Guerin & R. Carré, 357-366. Grenoble

House, D., G. Bruce, F. Lacerda & B. Lindblom. 1987a. 'Automatic prosodic analysis for Swedish speech recognition'. *Proc. European Conference on Speech Technology*, 215-218. Edinburgh.

House, D., G. Bruce, F. Lacerda & B. Lindblom. 1987b. 'Automatic prosodic analysis for Swedish speech recognition'. *Working Papers 31*, 87-101. Dept. of Linguistics, Lund University.

House, D., G. Bruce, L. Eriksson & F. Lacerda. 1988. 'Recognition of prosodic categories in Swedish: rule implementation'. *Working Papers 33*, 153-161. Dept. of Linguistics, Lund University.

House, D., G. Bruce, L. Eriksson & F. Lacerda. 1989. 'Prosodic parsing for Swedish speech recognition'. *Speech Transmission Laboratory - Quarterly Progress and Status Report (STL-QPSR) 1989:1*, 137-140. Dept of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm.

Lea, W. 1980. 'Prosodic aids to speech recognition'. *Trends in speech recognition,* ed. W. Lea, 166-205. Englewood Cliffs, N.J.: Prentice-Hall.

Mertens, P. 1987. 'Automatic segmentation of speech into syllables'. *Proc. European Conference on Speech Technology*, Edinburgh 1987, 9-12.

Mermelstein, P. 1975. 'Automatic segmentation of speech into syllabic units'. *Journal of the Acoustical Society of America* 58:4, 880-883.

Robins, R.H. 1966. *General linguistics. An introductory survey*. Bloomington: Indiana University Press.