# BLIND TECHNIQUE FOR AUTOMATIC SPEECH SEGMENTATION BASED ON SMMT METHOD

**Kamil Ekštein, Roman Mouček**

*Dept. of Computer Science and Engineering,*
*Faculty of Applied Science, University of West Bohemia,*
*Univerzitní 22, 306 00 Plzeň, Czech Republic*

**Abstract.** This paper presents an alternative approach to blind segmentation of speech signal based on newly designed *Spectral Moment Movement Tracking (SMMT)* method. The method was proposed during development stage of LASER (LICS Automatic Speech Evaluation/Recognition) speech recognition system which is being developed at Laboratory of Intelligent Communication Systems as a part of complex dialogue information system CIC (City Information Centre).

SMMT-based speech segmentation needs no phonetic modelling, no apriori information about incoming signal, nor training of any kind. The method comes out of few theoretical considerations described later in this paper, and thorough observation, modelling and analysis of speech signals. The method offers stable performance, efficiency, and simplicity of executive algorithms. Moreover, for Czech utterances used during testing it significantly overperformed "classic" segmentation methods including Bayesian Changepoint Detectors.

**Keywords:** Signal processing, automatic speech recognition, speech segmentation.

# 1. INTRODUCTION

Speech signal segmentation is one of the most important tasks in automatic speech recognition (ASR) process. The desiderative effect of signal segmentation is to split the signal in time into sections (termed *segments*) which correspond to some predefined phonetic units (phonemes, PLUs[1], alophones, etc.) and which can thus be analysed by acoustic-phonetic analysis methods. It was already shown that acoustic-phonetic analysis provides significantly better results when applied on well-segmented data—knowledge about location and duration of phonetic units in the incoming speech can facilitate the whole task and extensively increase recognition accuracy.

[1]Phoneme-Like Units

Yet there are tasks where suitable segmentation of some kind is absolutely necessary—for instance fluent speech recognition.

The purpose of the segmentation methods is therefore to estimate location of phonetic unit bounds as precisely as possible to match the reality. During the decades of ASR development, many methods were proposed. Nowadays, vast majority of the so called state-of-the-art speech recognizers benefit from *Hidden Markov Model (HMM)*-based phonetic decoding mechanisms. These mechanisms themselves have the signal segmentation task partially accomplished by their nature—the signal segmentation is closely connected with acoustic-phonetic decoding process here. Unfortunately the mentioned scheme, however, significantly decreases phonetic decoder efficiency because the HMM-based paradigm somewhat contradicts the possibility of thorough multimodal[2] analysis of the speech signal segment which is to be identified and assigned to a predefined phonetic class.

Although HMM-based ASR systems seem to offer respectable performance, they work surprisingly bad in adverse[3] and unexpected conditions and this together with the above stated facts directly encourages construction of speech recognizers using more analytical scheme. Such a scheme, however, is based on proper signal segmentation and thus justifies the search for it.

The other blind segmentation methods, i.e. those that need no training nor apriori knowledge, are mostly very complicated and demand substantial computational power. For example, lately popular *Bayesian Changepoint Detector*-based segmentation technique (described in details in [2]) requires at least 7 matrix multiplications, and an inverse matrix and determinant computation apart from other things to obtain probability of segment bound location. The depicted situation leads to need for a segmentation method which is both simple and efficient and it was also the main motive for the development of a new segmentation technique.

The newly proposed and designed SMMT method for speech segmentation estimates the phonetic unit bounds from determinate, easy-to-express changes in speech signal. The method is in fact generalization and simplification of SPMT (Spectral Peak Movement Tracking) method described in theory in [1].

# 2. THEORETIC ANALYSIS

Speech is extremely complex anthropic signal in which a linguistic message is coded by means of scheme that is not yet fully understood. In accordance with fundamental statements of information theory the signal varies in time to enable information transmission. Assuming that the effective information (required to set the segment bounds) is contained in signal spectrum, we need such a value whose time-domain progress expresses somehow the change in the analysed signal.

Such a value might be for example a distance between two consecutive spectra. But unfortunately two spectra with exactly the same mutual distance to reference one can be entirely different from phonetic point of view. Therefore a distance is not suitable for this task. Observing various speech spectrograms and analysing them, it was finally proposed to use moments (or better their time-domain traces).

[2]Here we mean several analysis methods applied on the same data, e.g. spectral analysis, time-domain structural analysis, homomorphic analysis, etc.

[3]Especially lack of training data or its poor quality severely degrades performance of HMM-based recognizers.

**Basic idea of Spectral Moment Movement Tracking (SMMT) group of methods is time-domain tracking of spectrum moment positions** assuming that moments sufficiently characterise their time series—here spectra of incoming speech signal. The moments are not understood in pure statistic way: The first general moment $\mu_1'$, i.e. mean value was used in it modified form having physical sense of "gravity centre" of given spectrum. Also the second moment $\mu_2$, i.e. variance $\sigma^2$, can be computed and thus enabling to model speech as Gaussian probability density distribution evolving in time domain with all possible consequences.

*Tracking* in SMMT technique context means mathematical analysis of the moment position change recorded in O(time×position) Cartesian system. The time-domain trace reflects very well the information carried by spectra of signal microsegments (see figure 2). Again, according to fundamental knowledge of information theory, local maxima of first-order derivative (computed by means of first-order difference) of moment position should indicate points of "information aggregation". Consequent consideration takes these points as phonetic information centres and thus makes speech segmentation possible. Phonemes (or better phoneme-like units) are expected to lie in between the mentioned points.

Another problem to solve is how to find these points. Simple derivative undoubtedly makes it possible to find local maxima of moment position change but does not enable to quantify the change easily. It is because the derivative produce another time series with "hills" and "valleys" and the moment position change quantification is here in fact probing the steepness of local maximum locality. More advantageous approach is to construct a tangent in each point of the original series and compare direction angle of the tangent with a predefined threshold value.

It is rather self-evident that the performance of SMMT-based speech segmenation is dependent on utterance language and would be for sure less efficient for well-tied languages like e.g. English. Fortunately Slavonic languages (where the application is presumed) are little tied and have very significant segmental structure facilitating SMMT technique enforcement.

# 3. EXECUTORY METHOD DESCRIPTION

*Spectral Moment Movement Tracking*-based segmentation of speech signal works as a simple *SISO (Single Input/Single Output)* system: The input data is a sequence of signal spectra, i.e. vectors containing values which correspond to relative signal energy in the appropriate narrow band. The output data is a series of estimated positions of the segment bounds.

Actually, *PSD (Power Spectral Density)* estimate vectors are used as an input and they can (and should) be interpolated and smoothed beforehand in order to obtain better results. Each PSD estimate vector is computed using FFT from 16 msec long (256 samples[4]) microsegment of the incoming speech signal. In pilot SMMT-based segmenter/labeler of which the results are presented later, there was used an extra processing (as mentioned above): Each PSD vector was linearly interpolated to get closer to double frequency resolution. Then the vector was smoothed using either a) simple averaging over several (3, 5 or 7) neighbouring values or b) substitution of the touched value by Gaussian-weighted linear combination of its

[4]As far as sampling rate 16 kHz is de facto technology standard in ASR, there is no reason why not use it as well :)

neighbours. Finally the vector components are compared with an task-specific[5] adjustable threshold value and if the component is smaller than the threshold it is set to zero not to influence the $\mu'_1$ calculation.

Such resulting (processed) vector is hereafter called *parametric vector* according to conventions of signal processing theory.

For each parametric vector acquired via the previously depicted process, a modified first general moment, $\mu'_1$, with physical meaning of "gravity centre", is computed:

$$\mu'_1 = \frac{\sum_{i=1}^{n} i \cdot h_i}{\sum_{i=1}^{n} h_i},$$

where $n$ is number of parametric vector components, and $h_i$ is $i$-th component value. It can be seen what is meant to be the modification: The statistical sense of the formula numerator is inverted to express a physical quantity. The vector values are taken as weights and the position indices as weighted values which enables to determine the position of "gravity centre" of the parametric vector.

In case a variance or dispersion, $\sigma^2$,

$$\sigma^2 = \overline{h^2} - (\overline{h})^2.$$

is also calculated for the parametric vector, it can be modeled by two moments as Gaussian probability distribution curve:

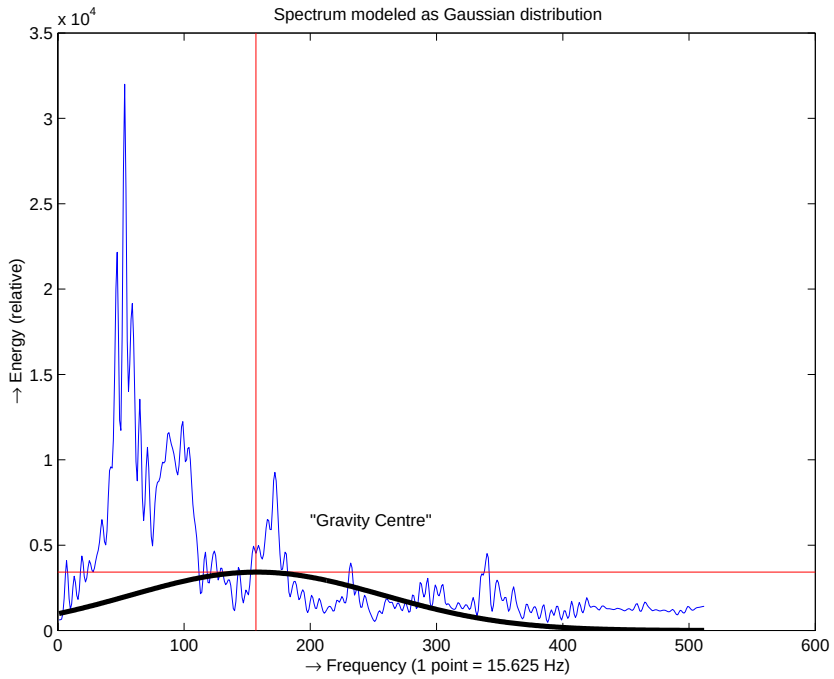


Figure 1: PSD estimate with "gravity centre" $\mu'_1$ and corresponding Gaussian curve

The above described computation transforms the whole input signal into a time-domain series of "gravity centre" $\mu'_1$ positions. The figure below shows a spectrogram

[5]Threshold values were determined experimentally during pilot application testing: It is slightly different for high quality speech and for, say, telephone transmitted utterance. The threshold setup also partially functions as noise suppression.

of processed utterance [h\ I= e d e j] (Czech word for "seek") with highlighted $\mu_1'$ positions for each microsegment[6]:

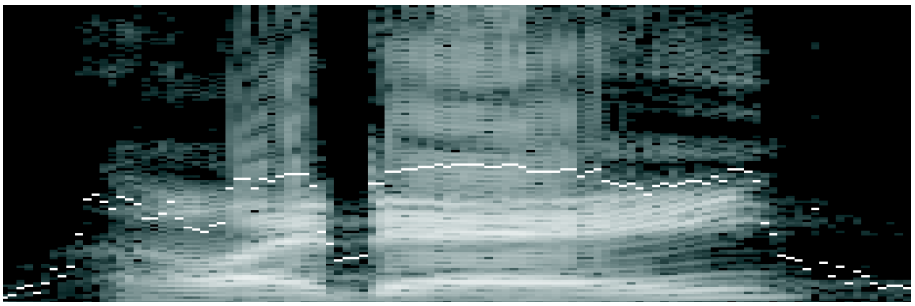

Figure 2: Spectrogram with highlighted $\mu_1'$ moment positions

The desired segment bounds are estimated according to first derivative of the obtained $\mu_1'$ time-domain series in the following manner: For each point, i.e. each signal microsegment, tangent of $\mu_1'$ series is computed using few neighbouring points[7] (3 or 5). If the tangent declination exceeds given limit, a new segment bound mark is placed at the touched point, i.e. right at the end of the corresponding microsegment.

# 4. EXPERIMENTAL RESULTS

SMMT-based speech segmentation was incorporated into pilot application—SpART-1 Tagger/Labeller auxiliary ASR software—to automatize speech signal labeling and was thoroughly tested on hundreds of Czech utterances of various quality. The overall score of SMMT-based method was approx. 73 % of correctly placed segment bounds while correctly placed means that the distance between the designed bound mark and ideal bound position was such that it does not influence phonetic value of bounded segment.

Generally the SMMT-based segmentation algorithm tends to place more segment bound marks than there actually is but those that match the real ones are usually placed absolutely accurately. In fact the "overdose" of bound marks fortunately does not degrade the performance of the ASR systems built on basis of classic HMM paradigm, hybrid ANN/HMM paradigm, or purely ANN-based one. A prospective error resulting from bound mark glut can be eliminated by some time aligning algorithm or just vocabulary search.

As an example let us examine the already mentioned [h\ I= e d e j] utterance in details. The utterance was segmented manually and automatically using SMMT—the table below summarises the experiment results and shows the bound mark positions:

[6]The image was grabbed from pilot application, SpART-1 Tagger/Labeller.

[7]Again, this is task-specific adjustable parameter because the method needs a bit different setup for each type of incoming signal.

| Phoneme | Factual start | Designed start (SMMT) |
|---|---|---|
| h\ | 0 | 0 |
| | | *(1280)*♣ |
| l= | 2704 | 2176 |
| e | 3792 | 3712 |
| d | 5056 | 4992 |
| e | 6224 | 5888 |
| j | 9664 | 9344 |
| | | *(10240)*♣ |
| | | *(11904)*♣ |

♣ These are superfluous bound marks placed at points where there are no phonetic bounds. Their occurrence is partly affected by setup of the method parameters.

The following figure enables direct comparison of manually and SMMT segmented speech signal as it shows two time-domain plots of the same utterance with bound marks set by hand (upper) and by SMMT-based algorithm (lower):

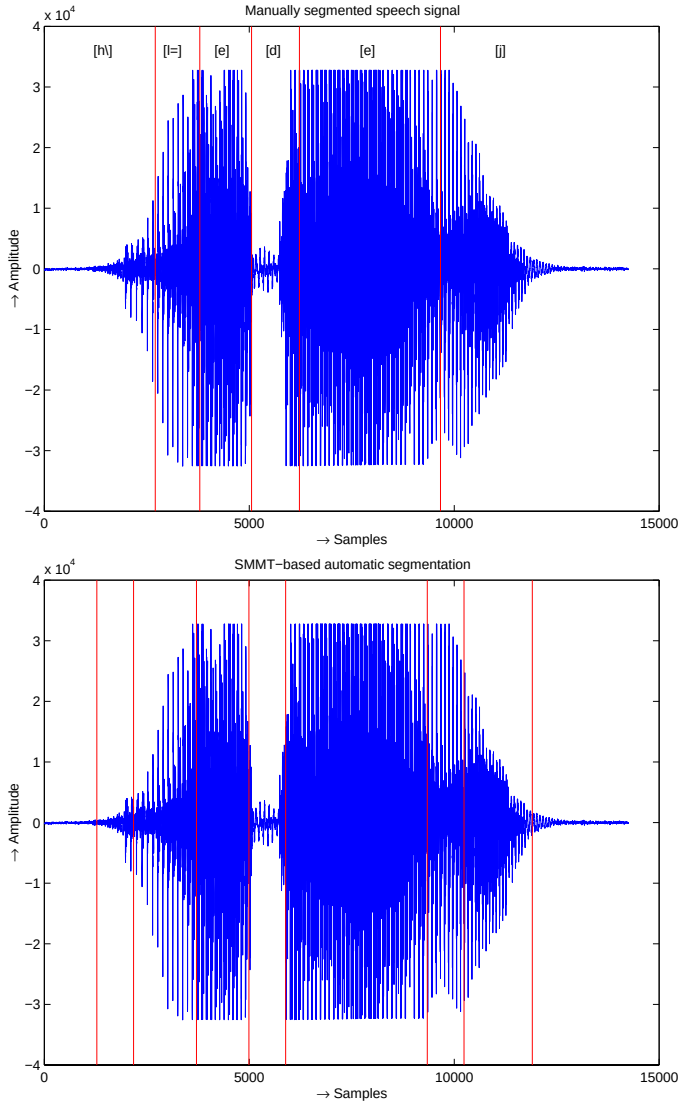


Figure 3: Segmented utterance with bound marks set manually (upper) and automatically (lower).

Phonetic analysis carried out consecutively proved that the segmentation of the utterance provided by SMMT-based method is well applicable. The redundant marks constitute no serious problem as long as they belong to the same phonetic class. It implies that a recognizer benefiting from SMMT-based segmentation algorithm would produce two concatenated identical phonetic transcription symbols instead of one in the worst case (here [h\ h\ I= e d e i j]). The mentioned situation is extremely easy to untangle.

# 5. CONCLUSION

The newly designed SMMT-based segmentation method was implemented in a pilot application within the frame of LASER speech recognition system—a speech signal labeling software. The method was tested on hundreds[8] of utterances of different duration, quality and, of course, spoken by different speakers.

The SMMT-based method proved fast and efficient during the tests. The overall score shows that it can surpass current segmentation methods while it is very lightweight from both implementational and computational point of view. The method was so far tested only for Czech utterances. It it highly probable that for such well-tied languages as for example English is, the performance would decrease.

# 6. ACKNOWLEDGEMENT

This research is in part supported by Grant of Czech Republic Ministry of Education, MSM 235200005 Information Systems and Technologies.

# REFERENCES

1. Ekštein, K.: "SpART 1—An Experimental Automatic Speech Recognizer", *Proceedings of SPECOM 2001*, pp. 93-96, Moscow, Russia, October 2001. ISBN 5-85941-092-1.

2. Čmejla, R., Sovka, P.: "Estimation Of Boundaries Between Speech Units Using Bayesian Changepoint Detectors", *Proceedings of TSD 2001*, pp. 291-298, Železná Ruda, Czech Republic, September 2001. ISBN 3-540-42557-8.

[8]The accurate number was not ascertained because many tests were performed by student research groups.