SPEECH SEGMENTATION USING BAYESIAN AUTOREGRESSIVE CHANGEPOINT DETECTOR

Roman ČMEJLA, Pavel SOVKA
Dept. of Circuit Theory
Faculty of Electrical Engineering
Czech Technical Technical University
Technická 2, Prague
Czech Republic
E-mail: cmejla@feld.cvut.cz

Abstract

This submission is devoted to the study of the Bayesian autoregressive changepoint detector (BCD) and its use for speech segmentation. Results of the detector application to autoregressive signals as well as to real speech are given. BCD basic properties are described and discussed. The novel two-step algorithm consisting of cepstral analysis and BCD for automatic speech segmentation is suggested.

Keywords

speech segmentation, sub-word boundaries, cepstral analysis, Bayesian methods, changepoint detector, autocorrelation

1. Introduction

Speech segmentation has been extensively studied because of the frequent application to speech analysis, coding, recognition, and speaker verification. In all these areas there is a need to describe the internal structure of speech including its changes in formant structure, voicing, pitch period, etc. A great number of methods based on different characteristics has been published. For example, correlation, zero-crossing, spectrum, cepstrum, autoregressive modeling (AR), wavelets and filter banks, hidden Markov models, neural networks, together with maximum a posteriori (MAP) estimation are often used. An overview together with some basic features of approaches mentioned before can be found in [1], [2], [3], [4], [5].

2. Bayesian changepoint detector

The autoregressive modelling is very often used for speech processing because of its simplicity and efficiency. That is why the MAP detector based on the general linear model [7], [6] was chosen.

Assume the speech can be modelled by the piecewise autoregressive model with abrupt changes in the order and coefficients (in other words: two different AR processes form the signal x[n])

$$x[n] = \begin{cases} \sum_{k=1}^{M_1} a_k x[n-k] + e[n], \; n<m \\ \sum_{k=1}^{M_2} b_k x[n-k] + e[n], \; n\geq m \end{cases}, \qquad (1)$$

where $x[n]$ stands for signal, $e[n]$ is zero mean noise, m is the time index of the parameters change, and a_k and b_k are parameters of models with orders M_1 and M_2.

This equation can be rewritten in the matrix form

$$\underline{d} = \underline{G} \cdot \underline{b} + \underline{e}, \qquad (2)$$

where $\underline{d}$ is the column data vector of size N with elements $x[n]$, $\underline{b}$ is the vector of model parameters with elements a_k and b_k, $\underline{G}$ is the $N \times M$ $(M = M_1 + M_2)$ matrix

$$\begin{bmatrix} x[1] \\ x[2] \\ \vdots \\ \vdots \\ x[m] \\ \vdots \\ \vdots \\ x[N] \end{bmatrix} = \begin{bmatrix} x[0] & x[-1] & \cdots & 0 & 0 & \cdots \\ x[1] & x[0] & \cdots & 0 & 0 & \cdots \\ x[2] & x[1] & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ x[m-1] & x[m-2] & \cdots & 0 & 0 & \cdots \\ 0 & 0 & \cdots & x[m] & x[m-1] & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ 0 & 0 & \cdots & x[N-1] & x[N-2] & \cdots \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ \vdots \\ a_{M1} \\ b_1 \\ \vdots \\ b_{M2} \end{bmatrix} + \begin{bmatrix} e[1] \\ e[2] \\ \vdots \\ \vdots \\ e[m] \\ \vdots \\ \vdots \\ e[N] \end{bmatrix}$$

By the marginalization of the likelihood function of $e[n]$, excluding nuisance parameters a_k, b_k and by the maximalization of likelihood function the result can be found in the form [7]

$$p(\{m\}|\underline{d}) \approx \frac{\left[\underline{d}^T \underline{d} - \underline{d}^T \underline{G} \left(\underline{G}^T \underline{G}\right)^{-1} \underline{G}^T \underline{d}\right]^{\frac{-(N-M)}{2}}}{\sqrt{\det\left(\underline{G}^T \underline{G}\right)}} \,. \qquad (3)$$

By searching for a maximum of this posterior density the position m can be found. It is necessary to point out that this method is very sensitive to changes in signal parameters and for a given data segment one maximum is always present. That is the reason why equation (3) can not be directly applied to speech segmentation.

Other problem is to estimate the orders M_1 and M_2 of AR models given by (1). Seven different model order criteria have been compared (final prediction error, minimal description length criterion, Bayesian Akaike information criterion, Schwartz's Bayesian criterion, phi criterion, and Akaike criterion) [10], [12], [13], [14], [16]. The Schwartz's Bayesian criterion (SBC) yield the best results. This evaluation is given by

$$SBC(M) = N \ln \hat{\sigma}_e^2 + M \ln N\,, \qquad (4)$$

where M is the number of parameters, N is the number of data and $\hat{\sigma}_e^2$ is the maximum likelihood estimation of the prediction (forecasting) error $\hat{\sigma}_e^2$.

2.1 Detector simulation with autoregressive signals

To learn more about the behaviour of BCD given by (3), the extensive simulation study was performed. The combination of many AR signals pairs with different orders and poles positions was generated. The detector was repeatedly applied to 100 or 1000 AR realisations and the maxima of the posterior density were evaluated using histograms. The results revealed the high sensitivity of detector to the errors in the AR orders estimation. Other crucial problem is the detector sensitivity to the length of the data vector. In this case some measures should be taken. The reasonable vector length lies between 60 and 1000 samples. Experiments confirmed that the detector precession is very high if its parameters are properly set up. The error in the determination of position *m* is less than several samples. More details can be found in [8].

2.2 Detector application to real speech signals

Due to the high sensitivity of BCD to spectral changes, as mentioned before, there is no possibility to use this detector directly to speech segmentation. To overcome this problem the preliminary segmentation using less sensitive methods than BCD must be used for the estimation of segmentation points between sound units. The BCD is then applied in the neighbourhood of these points to focus their positions. For the preliminary segmentation, the minima of the *sequence of autocorrelation function evaluated at the pitch period* (SACFL0) can be used. This approach is described in more details in [9]. The experimentation with this approach combining the autocorrelation function and BCD revealed that the detector often fails. In other words, the maximum of the posterior density is found either at the beginning or at the end of the data segment. The frequency of these failures is almost 50 %. The reason of this behaviour could be given by the fact that at the neighbourhood of each segmentation point the signal is highly nonstationary, and therefore the AR order can be often determined with a high inaccuracy. To solve this problem other approach must be applied. Instead of using segmentation points between sound units, the segmentation points inside stationary parts of signal should be used. In this case, the AR orders on the left and on the right sides of the data segment can be estimated with higher accuracy.

3. Speech segmentation

The whole automatic segmentation consists of three steps.

As discussed before, the first step of segmentation is to put a segmentation point to the centre of sound units composed of vowels and semivowels. Because of the simplicity and robustness, the approach suggested in [11] was applied and further modified. The key idea of this methods is based on the fact that vowels, in contradistinction of other parts of speech, have strong formants structure. Therefore the cepstral coefficients have large positive and negative values. In [11] the variance of 12 cepstral coefficients normalized by the energy was used for the segmentation. A slightly different approach consisting in the use of the *sum of squared covariance of cepstral coefficients* (SSCC) rather than the variance can be applied. After linear and nonlinear (median) filtering and thresholding the local maxima are searched for. Other modification of this approach uses the maxima of the SACFL0 instead of SSCC. In both cases, the nonlinear filtering yielding local maxima and minima called turning points [15] was applied alternatively to improve results.

In the second step BCD is applied between each pair of given segmentation points in stationary parts of speech. If the distance between these points overcomes thousand samples (depending on the sampling frequency used) the decimation must be used. But the decimation can cause a poorer detector performance. The result is always one point possibly lying between sound units. Because the left and right sides of the data segment are in stationary parts of speech the precision of segmentation is very high.

In the third step BCD uses the data between the stationary segmentation point and the segmentation point gained in the second iteration.

This approach is suitable for utterances with the vocal–consonant–vocal (VCV) structure.

4. Experiments and illustrations

The methods described above were evaluated on a small database consisting of isolated words as well as sentences of fluent speech. The sentences were spoken in different rates from five speakers. The whole number of vowels and semivowels *NV* was over 100 and the number of consonants *NC* was almost the same. For the evaluation of segmentation methods the criteria suggested in [11] were used: *insertions = in / NV missings = ms / NC*, where *ms* stands for the number of missing segmentation points in vowels and semivowels and *in* stands for the number of segmentation points inserted to consonants. The contents of used sentences were different. Therefore manually segmentation must be used for the evaluation of *in* and *ms* in this case. The final figure for both methods using SACFL0 or SSCC was similar to that one given in [11]; that is about 5% of insertions and 8% of missings.

The first segmentation step is illustrated in Fig. 1. The top part of this figure illustrates the whole sentence with resulting segmentation points. The content of this very quickly pronounced utterance is: "jaro už je tady sou tu slyšet hadi" [17]. The middle part of the figure depicts the SSCC, the bottom plot shows smoothed version of SSCC and estimated maxima exceeding the given threshold. One missing point for vowel "u" can be seen.

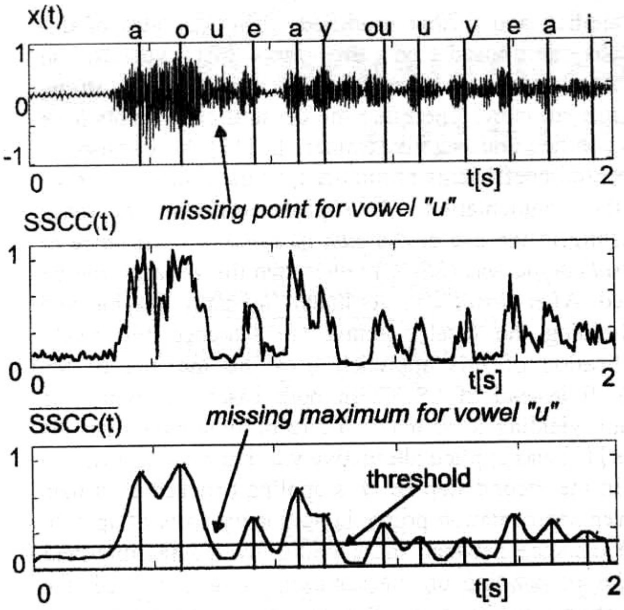

Fig.1 The first step of automatic segmentation: The evaluation of stationary segmentation points

Next three figures show examples of the segmentation using BCD. All three figures have the same structure. First plot is a signal (blue colour) with a segmentation point (blue line) gained by BCD and estimated AR orders. The red colour assignes the part of signal between two stationary segmentation points found in the first segmentation step. The second plot is a posterior density which is searched for a global maximum. The third plot shows a spectrogram together with a segmentation point gained by BCD (blue vertical line).

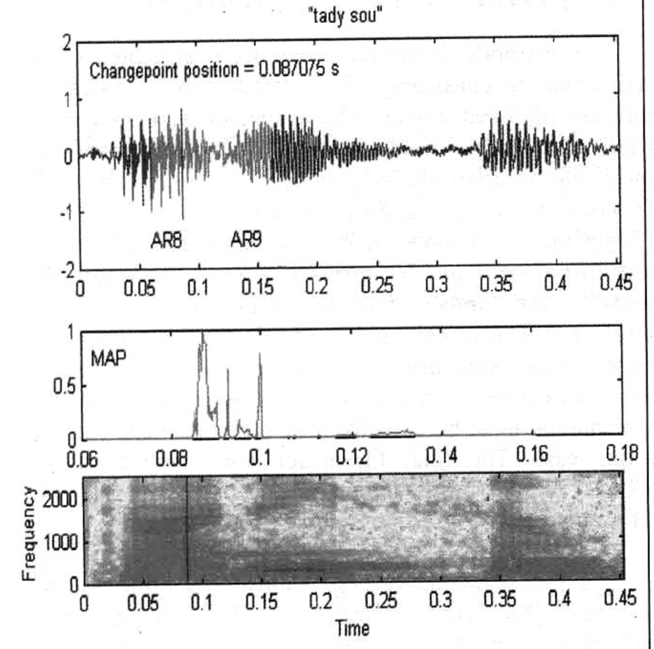

Fig.2 The second step of automatic segmentation using the Bayesian detector

Fig. 2 illustrates the second segmentation step when BCD is applied between two stationary segmentation points. The left point lies in the middle of vowel "a", while the right point is in the middle of vowel "y". In this case the results is the boundary between vowel "a" and the initial part of consonant "d". This boundary lies in the position, where the formant structure of "a" is strongly influenced by the following stop consonant "d" (the beginning of the transient).

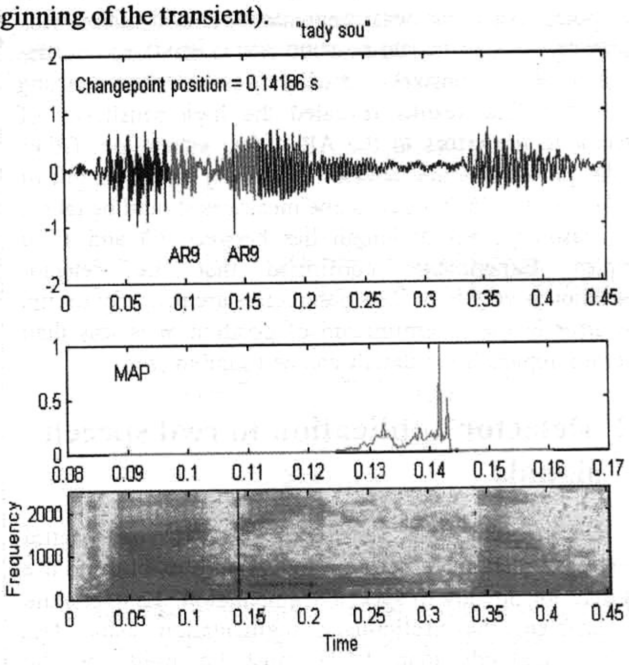

Fig.3 The third step of automatic segmentation using the Bayesian detector

Fig. 3 illustrates the third segmentation step when BCD is applied on the longer speech segment given by preceding segmentation procedure. As the result the boundary between the final part of consonant "d" and vowel "y" is found.

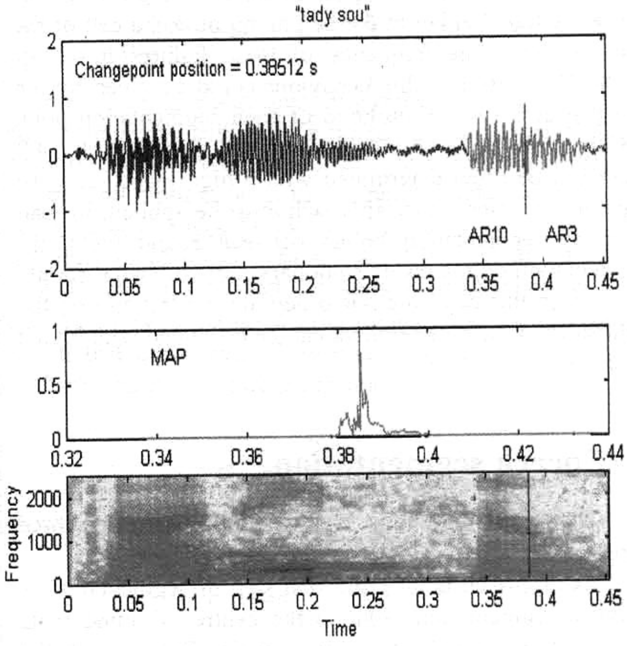

Fig.4 The example of phoneme boundary inside vowels

As it was mentioned BCD is very sensitive even to small changes in formant structure. This can be illustrated by Fig. 4, where the boundary between two vowels "o" and "u" is shown. On the other hand, BCD is not able to evaluate energy changes. Therefore BCD is not able to find start of occlusions because occlusions are determined by energy changes rather than formant changes.

The error rate of the whole segmentation method is given by the error rate in the first step. The error rate can be improved by another approach. The BCD is applied repeatedly with progressively growing data vector length. Results are evaluated by using the maxima in the histograms of posterior density as described in Section 2.1.

Acknowledgments

This research was supported by the grant GAČR 102/96/KO87 Theory and Application of Voice Communication in Czech.

The authors would like to express great thanks to Dr. W. J. Fitzgerald and Prof. P. J. W. Rayner for their valuable advices and support to Dr. R. Čmejla during his stay at the Engineering Department of Cambridge University.

References

[1] Rayner, P.J.W, Fitzgerald, W.J.: The Bayesian Approach to Signal Modelling and Classification. *In Proceedings of ECSAP'97*. The 1st European Conference on Signal Analysis and Prediction, Prague, Czech Republic, 1997, pp. 65-75.

[2] Li, T.H., Gibson, J.D.: Speech Analysis and Segmentation by Parametric Filtering. IEEE Trans.on Audio Processing, vol.4., no.3, May 1996, pp.203-213.

[3] Janer,L. at. all.: Wavelet Transforms for Non-Uniform Speech Recognition Systems. In *Proceedings ICSLP '96*, 4th Int. Conf. On Spoken Language, vol.4, pp. 2348-2351, Philadelphia, USA, 1996.

[4] Vostermans, A., Martens, J.P, Van Coile, B.: Automatic Segmentation and Labelling of Multi-lingual Speech. Speech Communication 19, 1996, pp. 271-193.

[5] Jeong, C.G., Hong Jeong., H.: Automatic Phone Segmentation and Labeling of Continuous Speech. Speech Communication 19, 1996, pp. 271-193.

[6] Procházka, A., Sláma, M., Pelikán, E..: Bayesian Estimators Use in Signal Processing, Neural Network World, vol. 6, No. 2, 1996, pp. 209-214.

[7] Ruanaidh, J.K.O, Fitzgerald, W. J..: Numerical Bayesian Methods Applied to Signal Processing. Springer-Verlag New York, Inc. 1996.

[8] Čmejla, R.:. Bayesian Detector: Experiments with Synthetic Speech Signals. Unpublished notes from the stay at University of Cambridge, United Kingdom, September, 1996.

[9] Čmejla, R., Sovka, P.:. System for Boundaries Detection and Sound Units Classification. Internal Research Report #R98-3, FEE CTU, Prague, Sept., 1998.

[10] Wei, W.W.S.: Time Series Analysis: Univariate and Multivariate Methods. Addison-Wesley Publishing Company, Inc., Wokingham, United Kingdom, 1995.

[11] Bán, L., Tatai, P.: Automatic Speech Segmentation for an Open Vocabulary Recognition System. *In Proceedings of ECSAP'97*. The 1st European Conference on Signal Analysis and Prediction, Prague, Czech Republic, 1997, pp. 303-306.

[12] Marple, S.L.: Digital Spectral Analysis with Applications, Prentice-Hall, Englewood Cliffs, New Jersey, 1987.

[13] Romberg, T.M., Black, J.L., Ledwige, T.J.: Signal Processing for Industrial Diagnostics, John Wiley, Chichester, 1996.

[14] Pukkila, M.T., Krishnaiah, P.R.: On the Use of Autoregressive Order Determination Criteria in Multivariate White Noise Tests, IEEE Trans.on ASSP, vol.36., no.9, September 1988, pp.1396-1403.

[15] Rychlik, I., Lindren G.: WAVE Analysis Toolbox a tutorial for use with Matlab 4.x. Manual ver. 1.1, Dep. Of mathematical statistics University of Lund, Box 118, S-22100 Lund, Sweden, 1995.

[16] Schloegel, A.: Time Series Analysis (TSA) Toolbox. http://www-dpmi.tu-graz.ac.at/~schloegl

[17] Psutka, J.: Komunikace s počítačem mluvenou řečí, Academia, Praha, 1995.

About authors...

Roman ČMEJLA was born in Louny, on April 25, 1962. He received M.S. degree in 1986 and Ph.D. degree in Communication Technology in 1993, both at the Faculty of Electrical Engineering of the Czech Technical University in Prague. He is currently working as assistant professor at Department of Circuit Theory. His research interests include adaptive methods of signal processing and knowledge-based approaches to automatic speech recognition.

Pavel SOVKA was born in Jihlava, Czechoslovakia, on February 4, 1957. He received the M.S. and Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering of the Czech Technical University (FEE CTU), Prague, in 1981 and 1986, respectively. From 1985 to 1991 he worked in the Institute of Radioengineering and Electronics of the Czech Academy of Sciences, Prague. In 1991 he joined the Department of Circuit Theory, FEE CTU. Since 1996 he has been an associated professor. Among his current research interests the application of adaptive systems to noise and echo cancellation, speech analysis, changepoint detection, signal separation can be found. Mr. Sovka is a member of the European Speech Communication Association (ESCA).