Detection of Acoustic Change-Points in Audio Streams and Signal Segmentation

Jindřich ŽĎÁNSKÝ

Dept. of Electronics and Signal Processing, Technical University of Liberec, Hálkova 6, 461 17 Liberec, Czech Republic

jindrich.zdansky@vslib.cz

Abstract. *This contribution proposes an efficient method for the detection of relevant changes in continuous stream of sound. The detected change-points can then serve for the segmentation of long audio recordings into shorter and more or less homogenous sections. First, we discuss the task of a single change-point detection using the Bayes decision theory. We show that it leads to a quite simple and computationally efficient solution based on the Bayesian Information Criterion. Next, we extend this approach to formulate the algorithm for the detection of multiple change-points. Finally, the proposed algorithm is applied for the segmentation of broadcast news audio-streams into parts belonging to different speakers or different acoustic conditions. Such segmentation is necessary as the first step in the automatic speech-to-text transcription of TV or radio news.*

Keywords

Audio stream processing, detection of acoustic changes, speaker segmentation, Bayesian information criterion, speech processing and recognition.

1. Introduction

Human voice is one of the essential means of communication and information exchange. Living in the information society at the beginning of the 3rd millennium, fast and accurate processing of information is the basis of success for businessmen, politicians, scientists and many others. Unfortunately, the recently available searching, sorting and information processing engines require the input data in textual form. Automatic processing of audio records of broadcast programs (namely news, debates or talks shows) with respect to their information content is still in its early stage. The most natural idea is to apply the existing full-text search machines on the transcribed version of those broadcast programs. Though, it assumes that first the spoken parts of the recordings are converted from speech to text.

In our lab (SpeechLab at the Technical University of Liberec) we have been involved in the long-term research dealing with the automatic processing of broadcast news. The currently developed system includes several parts, namely the signal processing unit, the acoustic segmentation unit and the speech recognition unit. In this paper we want to focus on the second one only, i.e. on the module that makes the segmentation of long audio streams into shorter parts that belong either to different speakers of different acoustic conditions (e.g. broad-band vs. narrow-band signals).

The first version of the system developed for the Czech broadcast news transcription was presented in [1]. The system processes the data at several levels. At the lowest one, the input waveform is parameterized and segmented into more or less acoustically homogenous parts (speaker turns, music parts, long silent or noise parts). At the higher level each segment is identified either as non-speech signal, which is skipped over, or as speech. The latter is sent to the speech recognition (and optionally also speaker recognition) module that provides the actual transcription.

In the original version described in [1], the stream segmentation module was designed to operate preferably in an on-line mode, i.e. directly on the continuously acquired signal. This approach allowed us to process the signal with only a small delay. On the other side, the accuracy and the reliability of the segmentation procedure was not so high. Therefore in this paper we propose an alternative method that is more suitable for off-line processing but gives significantly better results.

2. Single Change-Point Detection

The key idea behind the proposed method for the change-point detection task comes from the Bayes decision theory. That well-known theory is based on the assumption that the decision problem can be specified in probabilistic terms and that all of the relevant probability values are known or can be easily estimated.

2.1 Change-Point Detection Viewed as a Model Selection Task

The single change-point detection task can be viewed as a model selection problem. Let's suppose that we have

some data $D = (x_0,\ldots, x_N)$, which is a random sample from some unknown probability distribution for X. In addition, assume that the unknown probability distribution can be encoded by some statistical model with structure ω and parameters $\boldsymbol{\Phi}_\omega$. Now, suppose that our change detection problem consists of following model structures: single-state model for the case, when the data were produced by the single process/speaker, and the set of two-state models with respect to the change point location - for the two speaker/process case. See Fig. 1.

Given such a problem formulation we seek to find the model ω_i which best represents the data D. Using Bayes' *minimum-error-rate decision rule*, which is based on the maximum of the posterior probability $P(\omega_i|D)$, our problem can be re-introduced in the mathematical form as follows:

$$\omega_i = \arg\max_i P(\omega_i \mid D) \tag{1}$$

By applying the Bayes' rule and taking the logarithm we get:

$$\omega_i = \arg\max_i \log p(D \mid \omega_i) \tag{2}$$

where the prior probability $P(\omega_i)$ was omitted due to its *non-informative* nature. Now, our change detection task is simplified to the computing of log-likelihoods of the data D under the all possible models and choosing the one with the maximum value.

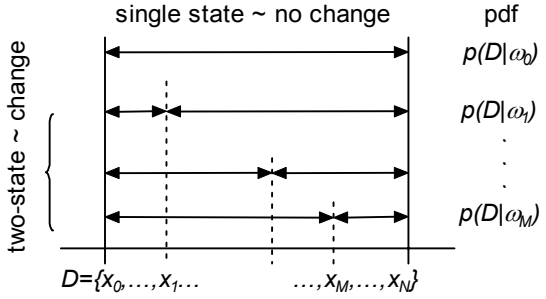


Fig. 1. Change point detection as a model selection problem.

2.2 Log-Likelihood Estimation

As was mentioned in the previous subsection, we assume that unknown probability distribution of X can be encoded by model structure ω and parameters $\boldsymbol{\Phi}_\omega$, but we are uncertain about both. Using the Bayesian approach we define discrete variable Ω whose possible values ω correspond to the possible model structures. In addition, for each model structure ω we define a continuous vector-valued variable Φ_ω, whose configurations $\boldsymbol{\Phi}_\omega$ correspond to the possible true parameters. Uncertainty about Φ_ω can be described by probability density function $p(\boldsymbol{\Phi}_\omega|\omega)$.

Given random sample D, the posterior distribution for each ω and $\boldsymbol{\Phi}_\omega$ can be expressed:

$$p(\boldsymbol{\phi}_\omega \mid D,\omega) = \frac{p(\boldsymbol{\phi}_\omega \mid \omega)\, p(D \mid \boldsymbol{\phi}_\omega,\omega)}{p(D \mid \omega)} \tag{3}$$

where

$$p(D \mid \omega) = \int p(\boldsymbol{\phi}_\omega \mid \omega)\, p(D \mid \boldsymbol{\phi}_\omega,\omega)\, d\boldsymbol{\phi}_\omega \tag{4}$$

is the *marginal likelihood* we are looking for.

In this paper we decided to employ *large-sample approximation* [2], [3] approach to estimate integral mentioned in the Equation 4. The basic idea behind large-sample approximations is that as the sample size N increases $p(\boldsymbol{\Phi}_\omega|\omega)\ p(D|\boldsymbol{\Phi}_\omega,\omega)$ can be approximated as a multivariate Gaussian distribution. This assumption leads to the *Laplace approximation* of the integral in the Equation 4:

$$\begin{aligned}\log p(D \mid \omega) &\approx \log p(D \mid \tilde{\boldsymbol{\phi}}_\omega,\omega) + \log p(\tilde{\boldsymbol{\phi}}_\omega \mid \omega) \\ &+ \frac{C}{2}\log(2\pi) - \frac{1}{2}\log|A|,\end{aligned} \tag{5}$$

where $C$is number of free parameters of the model ω, $\tilde{\boldsymbol{\Phi}}_\omega$ is *maximum a posteriori* (MAP) configuration of $\boldsymbol{\Phi}_\omega$ and A is the negative Hessian evaluated at $\tilde{\boldsymbol{\Phi}}_\omega$ of the function $\log(p(\boldsymbol{\Phi}_\omega|\omega)\ p(D|\boldsymbol{\Phi}_\omega,\omega))$.

Although this approach allows estimation of desired likelihood, computing Hessian is very numerically complicated. We can obtain another (computationally more efficient, but less accurate) approximation by retaining only those terms in Equation 5 that increase with N. Also for large N, $\tilde{\boldsymbol{\Phi}}_\omega$ can be approximated by the *maximum likelihood* (ML) configuration of $\hat{\boldsymbol{\Phi}}_\omega$, which results in so-called *Bayesian Information Criterion* (BIC):

$$\log p(D \mid \omega) \approx \log p(D \mid \hat{\boldsymbol{\phi}}_\omega,\omega) - \frac{\lambda}{2} C \log N\,. \tag{6}$$

The BIC approximation is quite intuitive. The first term measures how well the parameterized model predicts the data and the second term penalizes the complexity of the model. The importance of penalty term can be regulated by the penalty weight λ.

2.3 Efficient BIC Computation

In the consequent text we assume speech to be a multivariate Gaussian process in the *d-dimensional* cepstral space. The ML estimates for the Gaussian distribution are sample mean μ and sample covariance Σ. Under this assumption, BIC for single-state Gaussian model can be derived as follows:

$$\begin{aligned}BIC(\omega_0 \mid D) = &-\frac{N}{2}\log|\Sigma_0| - \frac{N}{2} - \frac{Nd}{2}\log(2\pi) \\ &- \frac{\lambda}{2}\left(d + \frac{1}{2}d(d+1)\right)\log N\end{aligned}\,. \tag{7}$$

BIC for the double-state Gaussian with respect to the change occurrence in the i-th frame can be formed as follows:

$$BIC(\omega_i \mid D) = -\frac{i}{2}\log|\Sigma_1^i| - \frac{N}{2} - \frac{Nd}{2}\log(2\pi) - \frac{N-i}{2}\log|\Sigma_{i+1}^N| - \lambda\left(d + \frac{1}{2}d(d+1)\right)\log N \quad (8)$$

where Σ^b_a denotes sample covariance computed from the data $x_a,\ldots,x_b$.

Hence, given the data sample D, the complete change detection process is simplified to the computation of covariance matrix Σ_0 and matrices Σ^i_1, Σ^N_{i+1}, for each frame $i = L,L+1,\ldots N-L$, where L is minimum number of frames to estimate covariance somehow reliably. Choosing the model ω_i with the highest *BIC* provides answer to the question where the change point is and whether there is any. Computation of matrices Σ^i_1, Σ^N_{i+1} can be done efficiently by forward, backward recursion, respectively.

3. Multiple Change Point Detection via Hierarchical Tree-Based Search

In this paper we describe hierarchical approach to the multiple change point retrieval. This approach is advantageous, because it enables to employ aforesaid single change point detection theory.

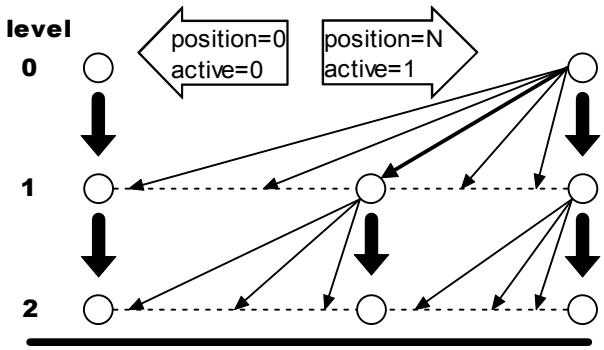
3.1 The Segmentation Algorithm

The idea behind the presented approach is that we reveal change points recursively, always looking for the only single change point, examining the area between two previously detected ones.

For each change point we define *node*, which has assigned two basic properties: time-*position* of the change in the data and the *active* property. The latter indicates whether the appropriate node is intended for further processing or not. Furthermore we define *arcs*, which denote computation of Equation 8. Choosing the *arc* with the highest value we get the most probable change point location. Putting Equations 7 and 8 in equality, we can computeλ for the most probable change point case. If the condition $\lambda \geq \lambda_{threshold}$ is fulfilled, a new node is established. The threshold $\lambda_{threshold}$ is the only free parameter of this algorithm, which needs to be estimated. Simplified algorithm outline and its graphical form are depicted in Fig. 2.

4. Performance Evaluation

The data used in our experiments have been collected as a part of pan-European Broadcast News Database by 6 institutions collaborating in the European COST278 action on Spoken Language Interaction in Telecommunication. Each participant prepared 3 hours of its national complete news broadcasts from public and/or private TV stations. The data set consists of wav files (16 kHz, 16 bit, mono),video files and transcription files. Each set is divided in two parts: two hours for the development and one hour for the testing purposes. More detailed description could be found in the paper [4]. Results presented in this paper were obtained from the Czech part of this database.




```
do
  for each node
    if is active
      assign arcs from position of node-1 to node
      get arc with maximum value
      compute λ
      if λ>threshold
        establish new node in the next level
      else
        active=0
       ---
     ---
    copy  current node to the next level
  ---
level++;
while there is any active node
```

Fig. 2. Simplified outline of the multiple change point detection algorithm.

Because most of the acoustic changes were speaker changes, we describe the speaker segmentation results. At first, computed boundaries are linked to the real ones if and only if computed boundary is the closest to the real one and vice versa. In addition, if the distance between them is smaller than 1 second, we call these boundaries *linked*. There are three statistical measures commonly used to acquire the segmentation accuracy:

$$recall = \frac{H}{N};\ precision = \frac{H}{H+I};\ F-rate = \frac{2*recall*precision}{recall + precision}, \quad (9)$$

where *N, H, I, D* are true, linked, inserted, deleted number of boundaries, respectively.

As was mentioned in the theoretical part of this document, we assumed speech to be multivariate Gaussian process in the cepstral space. Thus we converted input 16 kHz waveform into MFCC features, 12 coefficients computed every 10 ms from 25 ms window.

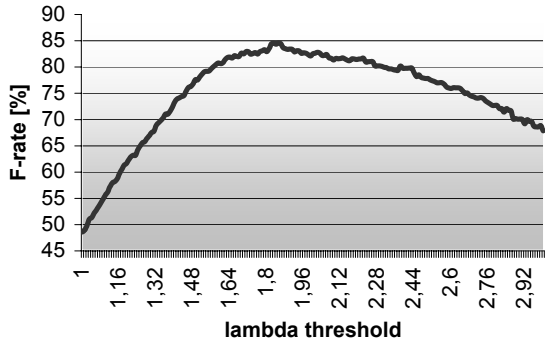


Fig. 3. Dependency of *F-rate* measure on the various penalty weights $\lambda_{threshold}$.

To estimate the optimal penalty weight $\lambda_{threshold}$, we decided to use *F-rate* measure as a criterion. Dependency of *F-rate* on the various penalty weights for the case of training data is shown in Fig. 3. Maximum *F-rate* value (84.61%) was obtained for the penalty weight $\lambda_{threshold} = 1.85$.

RECALL	PRECISION	F-RATE
87.81%	84.08%	85.91%

Tab. 1. Evaluation of the system performance - speaker segmentation results.

Evaluation of the segmentation system on the testing part of the database with penalty weight $\lambda_{threshold} = 1.85$ brought results summarized in Tab. 1. Proposed algorithm found 87.81% of all available speaker changes and 84.08% of produced changes were correct.

5. Conclusions

In this paper we have proposed an algorithm capable of segmenting long lasting sound records into shorter acoustically homogenous parts. The algorithm employs a hierarchical decision strategy that splits the considered part of signal into two parts at the point where the probability (measured by the BIC) of the potential change in signal characteristics is high enough.

The performance of this algorithm was evaluated on the speaker segmentation task within the broadcast news transcription system. The proposed method missed 12.19% of all existing speaker changes and 15.92% of the detected speaker changes were found as false alarms, which can be considered as good results in the quite complex task.

It should be noted that due to the assumptions used in the theory, the algorithm is not applicable for the detection of short segments. More exactly, it cannot detect segments shorter than 1 second, which is given by parameter L (see subsection 2.3), and the detection reliability is also disputable for segments shorter than 2 seconds. In practice, however, such extremely short speech segments are very rare in real broadcast recordings, and if they occur, they often carry a negligible piece of information.

The computational cost of this algorithm is small. When implemented on Pentium 2.4 GHz computers it does not takes more than 10% of the processor time.

Acknowledgements

This work has been supported by the Grant Agency of the Czech Republic (grant no. 102/05/0278).

References

[1] ZDANSKY, J., DAVID, P., NOUZA, J. An improved preprocessor for the automatic transcription of broadcast news audio stream. In *Proceedings of 8th International Conference on Spoken Language Processing ICSLP 2004*. JeJu (South Korea), 2004.

[2] KASS, R., RAFTERY, A. Bayes factors. *Journal of the American Statistical Association,* 1995, p. 773-795.

[3] KASS, R., TIERNEY, L., KADANE, J. Asymptotics in Bayesian computation. *Bayesian statistics 3*. Oxford University Press, 1988, pp. 261 - 278.

[4] VANDECATSEYE, A. et al. The COST278 pan-European broadcast news database. In *Proceedings of 4thInternational Conference on Language Resources and Evaluation LREC 2004*. Lisbon (Portugal), 2004.

[5] CHICKERING, D. M., HECKERMAN, D. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Technical report MSR-TR-96-08*. Microsoft Research, 1996.

About Authors...

Jindřich ŽĎÁNSKÝ was born in Česká Lípa in 1978. In 2002 he received master degree in electronics at the Czech Technical University (Faculty of Electrical Engineering) in Prague. In 2003 he joined the SpeechLab team at the Technical University of Liberec (TUL) as a PhD student. His research work is focused on robust speech recognition of Czech.