VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

DIARIZATION - WHO SPOKE WHEN

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE Bc. PETR ENT
AUTHOR

BRNO 2008

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

URČENÍ KDY KDO MLUVÍ

DIARIZATION - WHO SPOKE WHEN

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE AUTHOR	Bc. PETR ENT
VEDOUCÍ PRÁCE SUPERVISOR	Ing. PAVEL MATĚJKA

BRNO 2008

Abstrakt

Vyvíjený systém je postavený na standardních algoritmech pro segmentaci a clustering, využívajících Bayesovské informační kriterium. Pro detekci řeči je použit fonémový rozpoznávač. Systém je stále ve vývoji a zatím nebyl moc testován. V současnosti je zkoušeno pět různých řešení pro segmentaci a clustering. Dosud dosažené výsledky nebyly nijak porovnávány s jinými systémy, ale na první pohled vypadají slibně.

Klíčová slova

diarizace, segmentace mluvčích, BIC, Bayesovské informační kriterium

Abstract

Developed speaker diarization system is based on standard Bayesian Information Criterion segmentation and clustering algorithm. Phoneme recognizer is used for speech and non-speech detection. System is still not finished nor extensively tested, five possible solutions for segmentation and clustering are compared. No metric was used for obtained results, but at the first look they seem to be promising.

Keywords

diarization, speaker segmentation, BIC, Bayesian information criterion

Citace

Petr Ent: Diarization - Who Spoke When, diplomová práce, Brno, FIT VUT v Brně, 2008

Diarization - Who Spoke When

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Ing. Pavla Matějky

....
Petr Ent
January 10, 2008

Contents

Chapter 1

Introduction

The goal of speaker diarization systems is to segment an audio recording into regions according to the speaker. This task is sometimes also called ”Who Spoke When”. According to the NIST 2004 Rich Transcription evaluation [10], we consider this task to be done without a priori knowledge neither the number of speakers in audio recording nor their voices. Diarization is useful as a preprocessing phase in automatic speech to text transcription, where it can be used for speaker adaptation. It also may be used as a part of indexing information of audio archives. This work presents short introduction to speaker diarization systems based on Bayesian Information Criterion 3.1 and describes speaker diarization system which I have created under supervision of Ing. Pavel Matějka.

Chapter 2

State of the Art

Generally, the speaker diarization process could be divided into few basic steps, as is shown in the Fig. 3.2. The first one represent various kinds of preprocessing and speech activity detection. In next diarization stages, it is very important to work only with true speech data and not with background noise, music or silence. The second step is segmentation of speaker turn. The output of this stage is division of the utterance into the speaker homogenous segments. Third and final stage is assigning speaker identity to all segments. This is also the result of whole diarization process, to say how many speakers are in recording and to label the recording with currently speaking person's label.

Today's most commonly used principle in speaker diarization systems is Bayesian Information Criterion (BIC) 3.1. It is used in both stages: speaker segmentation and clustering. This technique was originally proposed by Chen and Gopalakrishnan [6]. Today's diarization systems usually besides BIC use some enhancements, such as speaker ID detection used by [13]. There are also some systems, which are not based on BIC at all and use different techniques, for example spectral clustering [9].

One of the systems based on BIC is the one from ICSI [11]. As it is typical to the most of today's systems, it works only with input data, with no need to prepare system before the diarization process starts (like training of models on outside data and so on). It can with advantage use more than one audio channel, if presented in audio recording. Wiener filter is used in preprocessing stage. Then, in case of more audio channels, they are mixed to one enhanced stream. Speech detection is done using three gaussian mixture models (model for speech, background noises and silence). Speaker segmentation stage is done via linear initialization as rough estimation and training and evaluation of models as fine tuning. Final stage uses agglomerative clustering in combination with BIC as a stopping criteria.

Similar approach is used in system from TNO [7]. The system works only with input data too. It uses only one channel from audio recording, even if more than one is presented. Speaker segmentation is done nearly same way as in ICSI system described above, the difference is that it uses only two models, the background noise model is omitted. This results to a little higher error in this stage for TNO's system. Speaker segmentation is done using Bayesian information criterion and is described in 3.2.2. Clustering stage is based on agglomerative clustering and Bayesian information criterion and is described in 3.2.3.

Chapter 3

Diarization using Bayesian Information Criterion (BIC)

3.1 Bayesian Information Criterion (BIC)

Bayesian Information Criterion (BIC) is statistical criterion for model selection [2]. It describes balance between goodness of fit of a given model to a data and number of parameters in model. In speaker diarization task, it is used for deciding if it is better for given data to use one big model or two separate smaller models. Equation for computing BIC is

$$BIC = log(X|M) - \frac{1}{2}\lambda N_m log(N_x) \tag{3.1}$$

where $log(X|M)$ is log-likelihood that model gives the data, N_m is number of parameters in model, N_x is the number of data samples and λ is penalty parameter.

In the speaker segmentation stage 3.2.2, BIC is used as is shown in Fig. 3.1. A model for the whole given speech signal with N_x samples is compared to two models A and B for a possible split. Model A consists of first N_a samples of given signal, model B consists of the rest N_b samples of the signal, assuming that $N_x = N_a + N_b$. If the ΔBIC in difference 3.2 is positive, it is better to model given speech signal with two smaller models, which usually means that two different speakers are in given data.

$$\Delta BIC = BIC_a + BIC_b - BIC_x \tag{3.2}$$

After substitution of BIC members, we have the following equation for BIC difference

$$\Delta BIC = \frac{1}{2}(N_x log|\Sigma| - N_a log|\Sigma_a| - N_b log|\Sigma_b| - \lambda N_m log(N_x)) \tag{3.3}$$

In theory, parameter λ should be equal to 1, and so it could be omitted. But previous works in this field [7] showed that parameter λ influences behavior of whole system a lot and could greatly differ from theoretical value (for example $\lambda = 14$ in TNO's system [7]) and thus it could not be omitted.

In clustering stage 3.2.3, BIC decides whether it is better to merge two segments together and consider them as from one speaker or let them split. The following equation is used

$$\Delta BIC = \frac{1}{2}(N_a log|\Sigma_a| + N_b log|\Sigma_b| + \lambda_c N_m log(N_x) - N_x log|\Sigma|) \tag{3.4}$$

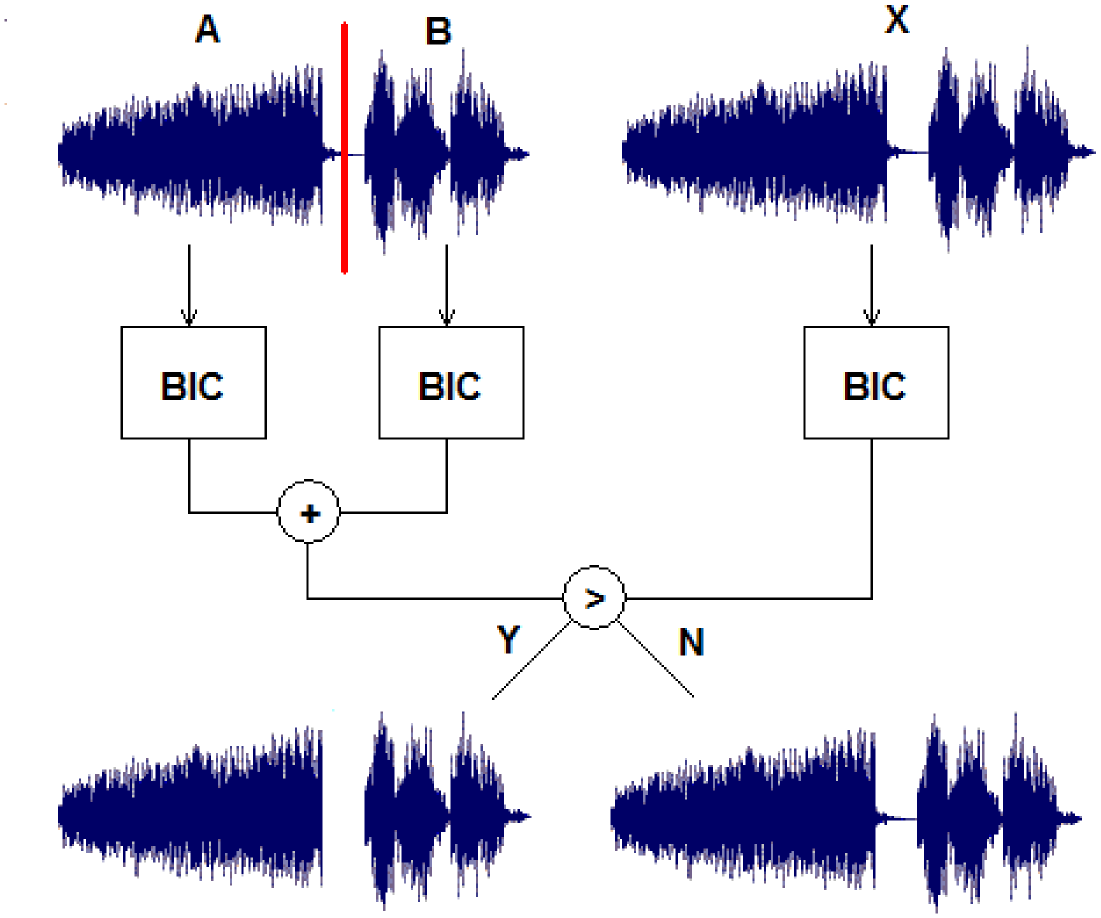

Figure 3.1: Use of BIC in segmentation stage.

where symbols N_x, N_a, N_b have same meaning as in equation 3.3. Note that equations differs only in sign and tunable penalty parameter λ_c.

Usage of equations 3.3 and 3.4 gives us two tunable penalty parameters λ and λ_c. They act as thresholds in splitting (λ) and clustering (λ_c) stages.

3.2 Steps in diarization process

As it was stated at the beginning 2, diarization process could be divided into three steps, as is seen in Fig. 3.2.

3.2.1 Preprocessing

The goal of this stage is to remove all the information which is not used later from audio stream. This is for example background noise, non-speech signals (music, noises like closing doors, steps, sneezing, crumpling of paper sheets and so on) and silence. If using more than one source of input data, such as multiple microphones, they are mixed together to one stream in this stage.

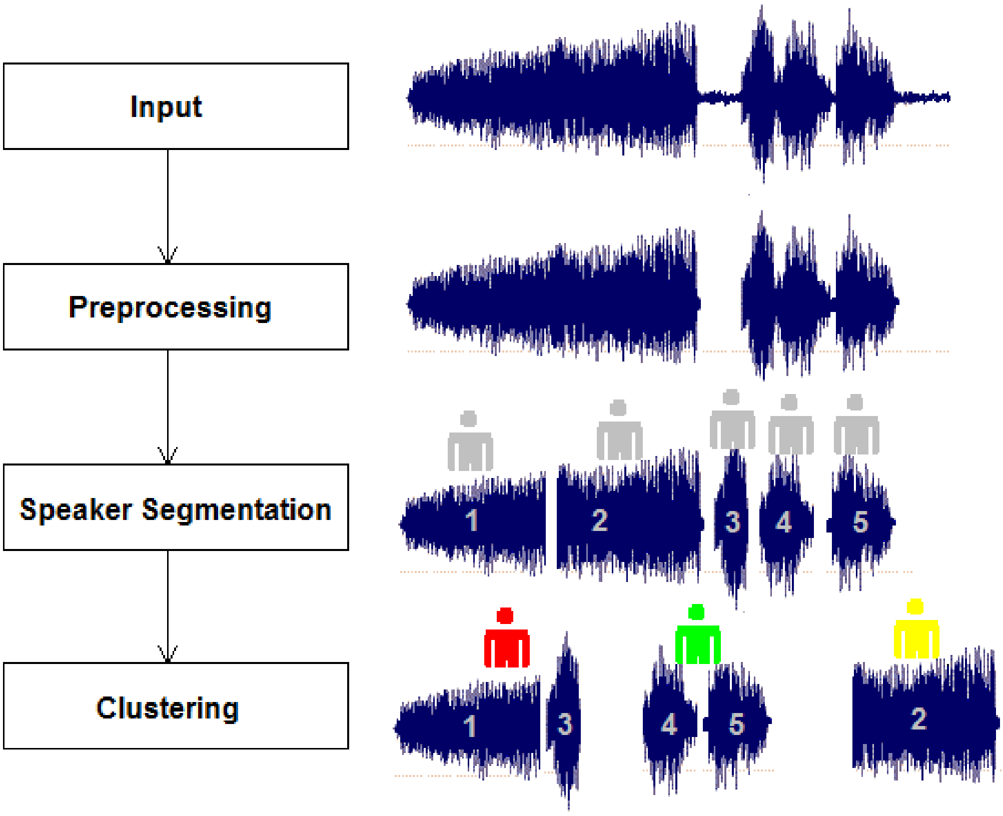

Figure 3.2: Steps and intermediate results in diarization process.

Background noise filtering could by done by various types of pass-band filters or Wiener filter. Several ways exist how to deal with problem of removing non-speech segments and silence parts. One of the most common is to use HMM models [4]. Three models are trained to classify speech music and silence. As a training data could be used either the processed audio record itself, or they could be trained separately before the diarization process starts. Only data classified as "speech" are used in next stages.

Other solution could be use of phoneme recognizer, for example phnrec [8]. Recognized phonemes are labeled as a speech, the rest of audio recording as silence and background noise. This approach is heavily dependent on the quality of phoneme recognizer. There has to be also some kind of post-processing, as the output of phoneme recognizer is usually large amount of very tiny speech segments.

3.2.2 Segmentation

The goal of segmentation stage is to split segments of speech from preprocessing stage into speaker homogenous parts. Again lot of possible approaches exist, but all of them have common base - BIC. Basically, we are always asking "Is it better to model this chunk of data by one model, or two models?". The things which are changing are when and how often we are asking and which data do we choose to be our "chunk".

One of the approaches, used by [7], is the following. A starting window of length t_w is taken from segment of speech, long enough to there can be expected possible speaker turn. Then a starting candidate breakpoint is chosen at time $t_c = t_p$. It divides window into two parts, A and B. ΔBIC is computed using equation 3.2. If it it positive, but smaller than value for previous candidate breakpoint, possible speaker turn is in this point and new segment is created from part A. Otherwise the algorithm follows with next candidate breakpoint $t_c = t_p + t_s, t_c = t_p + 2t_s ... t_c = t_w - t_p$. If the last candidate point is reached without speaker turn, the window is extended with another t_w time and the algorithm is run from the beginning with candidate breakpoints at $t_c = t_w - t_p, t_c = t_w - t_p + t_s ... t_c = 2t_w - t_p$. Graph of one speaker turn in segment is shown in Fig. 3.3.

Time t_p should be chosen carefully to avoid lack of training data for one of the models in edge parts of the window. Time t_s should be chosen short enough to not miss any speaker change, but long enough to avoid too much computational load.

Many variations of this method were presented, for example T^2 proposed by Zhou and Hansen [12], which chooses candidate breakpoints faster, and thus lowers computational load.

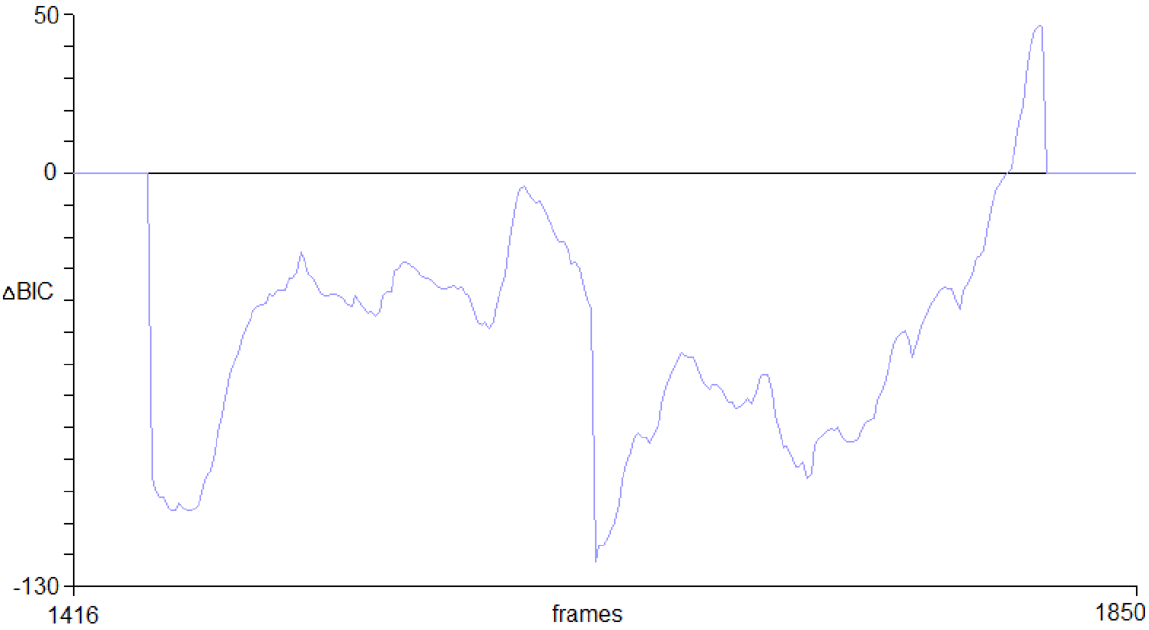

Figure 3.3: Graph of ΔBIC behavior. Axis y represents ΔBIC, axis x frames. Speaker turn is detected in frame 1823. Notice zero values from the beginning and at the end, those are the borders avoiding lack of training data, represented by t_p.

3.2.3 Clustering

Final stage of speaker diarization algorithm is merging segments from same speaker together. The technique most commonly used is called agglomerative clustering [3]. The process is very similar to splitting segments in segmentation part of diarization process 3.2.2, but it works in reversed way. It starts with small speaker homogenous segments, where every segment is considered as from different speaker. Then the algorithm chooses two the most similar segments using BIC 3.4 and if this score is positive, merge them. If no positive score is found between any two segments, the algorithm stops and each segment that remains is considered as one speaker. Diagram of the algorithm is shown in Fig. 3.4.

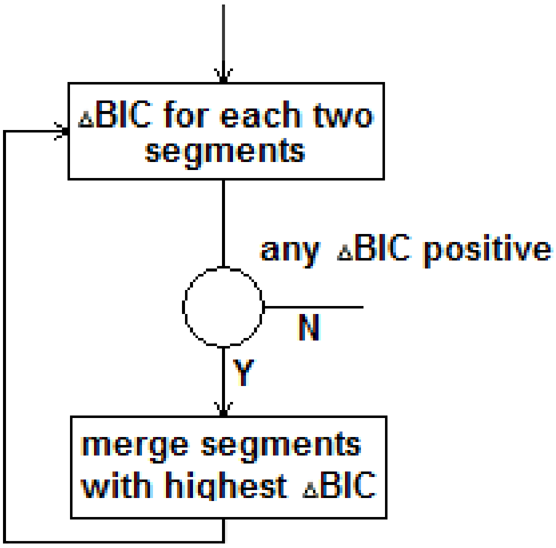

Figure 3.4: Diagram of clustering algorithm.

1. ΔBIC is computed for each two segments using equation 3.4
2. If any positive ΔBIC score is found, continue to step 3, else stop
3. Two segments with the highest ΔBIC are chosen to be merged, continue to step 1

Stoping criterion has to be tuned by λ_c parameter.

Chapter 4

Experimental diarization system

Our diarization system is based on BIC and is inspired mainly by the one from TNO [7]. No input filtering implemented yet. System is still in development and this stage is not crucial, if development data without background noise are chosen. However this is task for the future, if we want our system to be competitive. Speech, silence and background noise detection is performed via phoneme recognizer phnrec, developed in Speech@FIT group [8]. As it is mentioned earlier, output from phoneme recognizer is large amount of tiny segments. These segments have to be merged together. Currently, pause between two segments of speech shorter than t_{max} is considered as part of speech and these two segments and pause are merged together. Then all segments shorter than t_{min} are discarded, to avoid lack of data when training models. Used values t_{max} and t_{min} are in Tab. 4.1. [napsat hodnoy]

t_{max}	t_{min}
0.5s	10s

Table 4.1: Values t_{max} and t_{min}.

However this method is not without a flaw, short segments of speech are discarded and some non-speech data are used for training. Solution could be concatenation of all the speech segments from phoneme recognizer to one big, discarding only non-speech data. This brings task with mapping of merged segments back to the original audio recording, but it seems to be good direction for future improvements of our system.

Speaker segmentation phase uses algorithm described in 3.2.2. Variables used in algorithm are in Tab. 4.2.

t_w	t_p	t_s
5s	0.5s	0.1s

Table 4.2: Values t_w, t_p and t_s.

Possible improvement of this algorithm could be division into two passes, first pass will be with larger step t_s and determine rough position of speaker turn, second pass will be with smaller step and fine tune the location of speaker turn. Second pass will be performed only on close neighborhood of the point from first pass.

Clustering part of algorithm works as is described in 3.2.3. Possible improvements of this part will be oriented towards speed. This is the part where the most of time is spent, because there need to be computed ΔBIC for each pair of segments. This brings quadratic complexity. One of speedups could be dividing segments into few groups at the beginning, run clustering algorithm on each of them, then make new group from the results and run algorithm again on this group. This approach has still quadratic complexity, but with slightly better coefficients.

Chapter 5

Experimental setup

As the system is still in development phase and changes are made very often, we have not extensively tested it yet. Currently we use one 30 seconds long Czech audio recording for the purpose of tuning parameters and choosing the best components for each stage. When we finish this phase, we plan to test the best configuration on data from AMI project [1]. The AMI project data we have consist of about 3 hours of English spoken meetings. In further future, we plan to test our system on data from NIST evaluations [5], which consist of various meetings and various languages.

Chapter 6

Results

We have tested the following model setups:

- Model with diagonal covariance matrix
- Model with diagonal covariance matrix, map adaptation of means
- Model with diagonal covariance matrix, map adaptation of means and variances
- Model with full covariance matrix
- Model with full covariance matrix, map adaptation of means

Results are shown in Fig. 6.1. Ideal segmentation and diarization are in the first row, all tested setups follows. The same model was used in segmentation and clustering stage for each setup.

In the task of speaker segmentation, the best results came from the model with diagonal covariance matrix. This is a bit surprising, as we expected that more parameters in model with full covariance matrix would mean better results. It is possible that this result is not correct due to small amount of development data, so it will has to be proved by testing on more data. Results from other methods are approximately equal.

Results of clustering are very similar to those of speaker segmentation. Again, model with diagonal covariance matrix proves itself to be the best one. The worst results has full covariance matrix model with map adaptation. However this results are rather disputable. It is possible that bad segmentation could cause errors in clustering stage. More tests has to be made to prove this results.

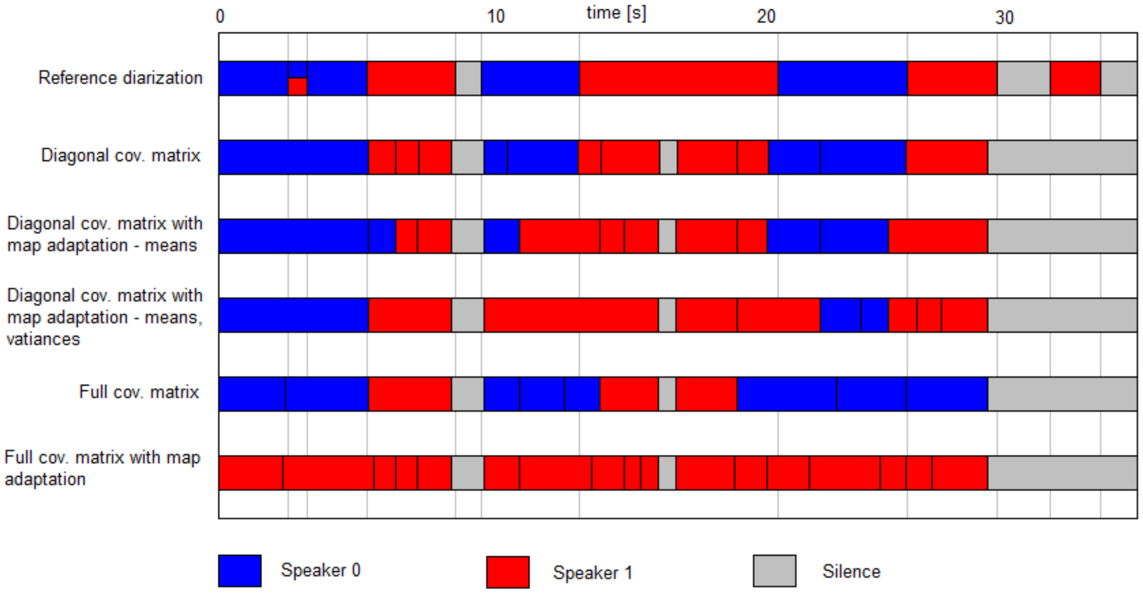

Figure 6.1: Results of tested methods. There are two speakers in audio recording. Gray lines mark ideal speaker segmentation, red and blue colors mark ideal clustering. Light gray space is silence.

Chapter 7

Conclusion

Simple, but functional diarization system was created. Next step will be testing of all proposed methods for each stage on larger data. Finally, the best combination of them will be put together and tested on data from NIST evaluations. At this point, it will be also possible to compare our system with others that took part in NIST evaluations.

There is also lot of space for improvements. The highest priority has better use of phoneme recognizer, more precise detection of speaker turn, noise filtering on input audio and speedup of whole system, as it was proposed in 4. In further future, possible improvement could be including speaker identification into system.

Bibliography

[1] Ami project. [online], http://www.amiproject.org. [rev. 2008-01-04], [cit. 2007-12-28].

[2] Bayesian information criterion. [online], http://en.wikipedia.org/wiki/Bayesian_Information_Criterion. [rev. 2008-01-04], [cit. 2007-01-02].

[3] Cluster analysis. [online], http://en.wikipedia.org/wiki/Cluster_analysis. [rev. 2008-01-04], [cit. 2007-01-03].

[4] Hidden markov models. [online], http://en.wikipedia.org/wiki/Hidden_Markov_model. [rev. 2008-01-04], [cit. 2007-01-03].

[5] National institute of standards and technology, speech group. [online], http://www.nist.gov/speech. [rev. 2008-01-04], [cit. 2007-12-22].

[6] S. Chen and P. Gopalakrishnan. *Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion.* in Proc. DARPA Broadcast News Transcription Understanding Workshop, 1998.

[7] D. van Leeuwen. *The TNO Speaker Diarization System for NIST RT05s Meeting Data.* NIST 2005 Spring Rich Transcription Evaluation Workshop, 2005.

[8] P. Matějka, L. Burget, P. Schwarz, and J. Černocký. Nist language recognition evaluation 2005. In *Proceedings of NIST LRE 2005*, Washington DC, US, 2006.

[9] H. Ning, M. Liu, H. Tang, and T. Huang. *A Spectral Clustering Approach to Speaker Diarization.* [online], http://www.interspeech2006.org/papersearch/IS06papers/IS061607.PDF. [rev. 2008-01-04], [cit. 2007-12-22].

[10] NIST. *Fall 2004 Rich Transcription (RT-04F) Evaluation Plan*, 2004.

[11] Ch. Wooters and M. Huijbregts. *The ICSI RT07's Speaker Diarization System.* [online], http://www.icsi.berkeley.edu/pubs/speech/ICSI_RT07_diarization.pdf. [rev. 2008-01-04], [cit. 2007-12-22].

[12] B. Zhou and J. Hansen. *Efficient Audio Stream Segmentation via the combined T^2 Statistic and Bayesian Information Criterion.* IEEE Transactions on Speech and Audio Processing, 2005.

[13] Ch. Zhu, C. Barras, S. Meignier, and J. L. Gauvain. *Improved Speaker Diarization using Speaker Identification.* [online], http://www.limsi.fr/Rapports/RS2005/chm/tlp/tlp1/index.html. [rev. 2008-01-04], [cit. 2007-12-22].