ITMO University
Department of speech information systems

Speaker Diarization System Based on Probability Linear Discriminant Analysis

technical report

Author:
Oleg Kudashev
kudashev@speechpro.com

Scientific adviser:
Timur Pekhovsky
tim@speechpro.com

Editor:
Olga Khomitsevich
khomitsevich@speechpro.com

Saint Petersburg
2015

Abstract

This report presents a brief overview of the results obtained in the thesis by O. Kudashev "Speaker Diarization System Based on Probability Linear Discriminant Analysis" [1]. The thesis deals with the speaker diarization task and proposes a novel method of solving it. It demonstrates that using Probability Linear Discriminant Analysis (PLDA) and the Total Variability (TV) approach makes it possible to solve the speaker diarization task with high accuracy even if the number of speakers is unknown.

Keywords: speaker diarization, speaker segmentation, clustering, probability linear discriminant analysis, factor analysis, speech recognition

1 Introduction

The speaker diarization task involves detecting and clustering speech segments of an audio stream so that all segments from one cluster belong to the same speaker, and vice versa. The solution of this task gives the answer to the question "Who spoke when?"

The problem under consideration is an essential part of many automatic speech processing systems, such as speech recognition and speaker recognition systems. Fast development of speech processing methods leads to increasingly demanding requirements for speaker diarization accuracy.

This work was inspired by the results obtained in [2–4] where it was shown that using a wide range of a priori information makes it possible to obtain state-of-the-art results for the speaker diarization task. The combination of Variational Bayesian Analysis (VBA) in conjunction with Joint Factor Analysis (JFA) leads to high accuracy in speaker diarization of phone conversations. However, the abovementioned papers deal with the case when the number of speakers is known a priori. That limits the range of applicability of a speaker diarization system based on such methods.

Some of the latest methods used in speaker recognition are based on Probability Linear Discriminant Analysis (PLDA) [5, 6]. These methods deal with the "i-vector" representation of speech segments in Total Variability (TV) space [7] and show good performance especially in the case of verification using short utterances [8, 9]. It is the author's opinion that PLDA is a convenient tool for data modeling and promises an effective solution of the problem at issue. In this thesis the author proposes applying PLDA in the speaker diarization field.

2 Speaker diarization system based on PLDA

2.1 Probability Linear Discriminant Analysis

Both PLDA and LDA deal with inter- and within-speaker variability of i-vectors. These methods estimate the directions in the TV space which maximize speaker discrimination. As distinct from LDA, PLDA operates within a probability framework and makes it possible to apply a wide range of corresponding methods for data analysis.

In accordance with the PLDA approach [6], a randomly selected i-vector is represented as follows:

$$w = s + c \,, \tag{1}$$

where w is an i-vector; s and c are statistically independent vectors of speaker and channel factors; w, s, c have Gaussian distribution.

Following the JFA, it is assumed that s and c can be presented as a linear combination of the hidden factors. Let $W = \{w_r \in \mathbb{R}^D\}_{r=1}^R$ be a set of i-vectors of the same speaker. Then each of these vectors is represented as follows:

$$w_r = m + Vy + Ux_r + \epsilon_r \,. \tag{2}$$

Here m is the mean vector; x,y are random vectors with the dimension N_x, N_y with a standard normal distribution; ϵ is a random vector with a Gaussian distribution with a zero mean and the precision matrix Λ.

Following [8], in this thesis a simplified model is considered in which the dimension N_x equals 0.

2.2 PLDA for the speaker diarization task

The key aspect of the PLDA approach is the estimation of the model parameters (m, V, Λ). In terms of speaker recognition task this estimation is made using the following training data of the i-vectors:

$$W = \bigcup_{s=1}^{S} \bigcup_{r=1}^{R(s)} w_{s,r} \,, \tag{3}$$

where S is the number of speakers; $R(s)$ is the number of sessions of the speaker s. Each $w_{s,r}$ is represented as follows:

$$w_{s,r} = m + Vy_s + \epsilon_{s,r} \,. \tag{4}$$

Obviously, the approach described above is unsuitable for speaker diarization task, since speaker diarization methods deal with very short

speech segments derived from the same audio recording. Thereby, a different model of PLDA should be applied. The author proposes to split each speaker session into several short speech segments. Thus, the following training data should be obtained:

$$W = \bigcup_{s=1}^{S} \bigcup_{r=1}^{R(s)} \bigcup_{k=1}^{K(r,s)} w_{s,r,k} \ , \tag{5}$$

where $K(r,s)$ is the number of short speech segments in the recording r of the speaker s. Each i-vector $w_{s,r,k}$ is represented as follows:

$$w_{s,r,k} = m + Vy_{s,r} + \epsilon_{s,r,k} \ . \tag{6}$$

The mean vector m in (6) can be assumed to be equivalent to zero, since it may be estimated and subtracted preliminarily from both training and evaluation data.

The main aspect of the equation (6) is that the speaker factors $y_{s,r}$ are fixed for the session s,r and not for the speaker s.

Firstly, the result is that the proposed approach imposes corresponding changes on the procedure of estimating PLDA model parameters. Such a model significantly reduces variability in estimating the speaker factors $y_{s,r}$ within an audio recording.

Secondly, the index s,r can be replaced by a new index which denotes the index number of the recording, so the parameter estimation does not need the information about the speaker identifier present in the recording.

2.3 Speech segment clustering

Let us briefly consider the problem of clustering a set of i-vectors estimated on the speech segments $W = \{w_1, \ldots, w_R\}$.

Let us suppose we have M of all possible ways $\{\Theta_m\}_{m=1}^{M}$ to cluster W. In this case the optimal solution of the clustering task $\widehat{\Theta}$ can be obtained using the Maximum-Likelihood criterion:

$$\widehat{\Theta} = \underset{m=1,\ldots,M}{\arg\max} P(W|\Theta_m) \ . \tag{7}$$

Unfortunately, M grows dramatically with the growth of the number of speech segments. This results in the impossibility of an exhaustive search for the optimal clustering $\widehat{\Theta}$ for the speaker diarization task.

The author proposes to divide the problem mentioned above into two sub-tasks:

1. The first of these sub-tasks consists of searching for optimal clustering solutions $\{\Theta_1, \ldots, \Theta_{K_{max}}\}$ with the known number of clusters $\{1, \ldots, K_{max}\}$.

2. Next, it is required to estimate the likelihood function $P(W|\Theta_m)$ for the relatively small set $\{\Theta_1, \ldots, \Theta_{K_{max}}\}$ and to obtain the resulting clustering using the equation (7).

2.3.1 Clustering for a known number of speakers

VBA-PLDA

Following [3], Variational Bayesian Analysis (VBA) in the context of the proposed PLDA model was used for this task (the **VBA-PLDA** method).

Let us consider the input set of i-vectors $W = \{w_i\}_{i=1}^R$ corresponding to R speech segments. Let $Y = \{y_k\}_{k=1}^K$ be the hidden vectors of the speaker factors. As in [3], let $Z = \{z_i\}_{i=1}^R$ be the hidden indicator vectors whose components are defined as follows: $z_{i,k} = 1$ if the speaker k is talking in the segment i and $z_{i,k} = 0$ otherwise. Let π_k be the a priori probability of the presence of the speaker k in a speech segment. The a priori distributions are expressed as:

$$P(Y) = \prod_{k=1}^{K} P(y_k) = \prod_{k=1}^{K} \mathcal{N}(y_k|0, I) \,, \tag{8}$$

$$P(Z) = \prod_{i=1}^{R}\prod_{k=1}^{K} P(z_{ik}) = \prod_{i=1}^{R}\prod_{k=1}^{K} \pi_k^{z_{ik}} \,. \tag{9}$$

Supposing that these distributions are independent:

$$P(Y, Z) = P(Y)P(Z) \,. \tag{10}$$

In this case the log likelihood function is given by:

$$\begin{aligned} \ln P(W, Y, Z|V, \Lambda) = &\sum_{i=1}^{R}\sum_{k=1}^{K} z_{ik}\left(-\frac{1}{2}(w_i - Vy_k)^T\Lambda(w_i - Vy_k)\right) + \\ &+ \sum_{k=1}^{K}\left(-\frac{1}{2}y_k^T I y_k\right) + \sum_{i=1}^{R}\sum_{k=1}^{K} z_{ik}\ln\pi_k + C \end{aligned} \,, \tag{11}$$

where V, Λ are the parameters of the PLDA model.

The goal of VBA is the search for an appropriate $Q(Y, Z)$ for the a posteriori distribution $P(Y, Z|X, V, \Lambda)$ of hidden vectors which maximizes

the low bound $\mathcal{L}(Y,Z)$:

$$\mathcal{L}(Q) = \int Q(Y,Z) \ln \frac{P(W,Y,Z|V,\Lambda)}{Q(Y,Z)} \, d\Theta \; . \tag{12}$$

According to VBA, the only factorization of $Q(Y,Z)$ is assumed to be:

$$Q(Y,Z) = Q(Y)Q(Z) \; . \tag{13}$$

Then the formulas for variation approximation of $Q(Y)$, $Q(Z)$ are:

$$\begin{aligned} \ln Q(Z) &= \langle \ln P(W,Y,Z|V,\Lambda) \rangle_Y + C \; , & (14) \\ \ln Q(Y) &= \langle \ln P(W,Y,Z|V,\Lambda) \rangle_Z + C \; . & (15) \end{aligned}$$

Since $P(Y)$ and $P(Z)$ are conjugate priors, $Q(Y)$ and $Q(Z)$ will be sought in the form:

$$\begin{aligned} Q(Z) &= \prod_{i=1}^{R} \prod_{k=1}^{K} q_{ik}^{z_{ik}} \; , & (16) \\ Q(Y) &= \prod_{k=1}^{K} \mathcal{N}(y_k | \hat{y}_k, \hat{\Sigma}_{y_k}) \; . & (17) \end{aligned}$$

Therefore:

$$\begin{aligned} \langle z_{ik} \rangle_Y &= q_{ik} \; , & (18) \\ \langle y_k \rangle_Y &= \hat{y}_k \; , & (19) \\ \left\langle y_k y_k^T \right\rangle_Y &= \hat{\Sigma}_{y_k} + \hat{y}_k \hat{y}_k^T \; . & (20) \end{aligned}$$

In summary, the formulas for variation approximation of $Q(Y)$, $Q(Z)$ are:

$$\ln Q(Y) = \sum_{k=1}^{K} \left(-\frac{1}{2} (y_k - \hat{y}_k)^T \hat{\Sigma}_k^{-1} (y_k - \hat{y}_k) \right) + C \; , \tag{21}$$

where

$$\begin{aligned} \hat{\Sigma}_k^{-1} &= (I + \sum_{i=1}^{R} q_{ik} V^T \Lambda V) \; , & (22) \\ \hat{y}_k &= \hat{\Sigma}_{y_k} V^T \Lambda \sum_{i=1}^{R} q_{ik} w_i \; . & (23) \end{aligned}$$

$$\ln Q(Z) = \sum_{i=1}^{R} \sum_{k=1}^{K} z_{ik} \ln \hat{q}_{ik} + C \ , \tag{24}$$

where

$$\ln \hat{q}_{ik} = -\frac{1}{2} w_i^T \Lambda w_i + \hat{y}_k^T V^T \Lambda w_i - \frac{1}{2} \operatorname{Tr}\left(V^T \Lambda V (\hat{\Sigma}_{y_k} + \hat{y}_k \hat{y}_k^T) \right) + \ln \pi_k \ . \tag{25}$$

The important question is how to determine the initial values of $q_{i,k}$. One way is random initialization. It can also be done by using results obtained from another clustering method.

KERNEL K-MEANS

An alternative method for i-vector clustering was also proposed. This method is based on K-means algorithms with a non-linear kernel [10] (the **KERNEL K-MEANS** method). The main idea is using the generalized K-means algorithm which uses a non-linear kernel defined by a symmetric similarity matrix. The elements of the similarity matrix $A_P = \{a_{i,j}^P\}_{i,j=1}^R$ are non-negative and defined as follows:

$$\begin{aligned} a_{i,j}^P &= P(H_s|w_i, w_j) = \\ &= \frac{P(w_i, w_j|H_s)P(H_s)}{P(w_i, w_j|H_s)P(H_s) + P(w_i|H_d)P(w_j|H_d)P(H_d)} \ , \end{aligned} \tag{26}$$

where H_s is the hypothesis that w_i and w_j belong to the same speaker; H_s is the hypothesis that w_i and w_j belong to different speakers.

According to [8], the equation (26) can be rewritten as follows:

$$a_{i,j}^P = \frac{e^{score(w_i,w_j)+\lambda}}{e^{score(w_i,w_j)+\lambda} + 1} \ , \tag{27}$$

where

$$\begin{aligned} score(w_i, w_j) =& \ln \mathcal{N}\left(\begin{bmatrix} w_i \\ w_j \end{bmatrix} | \begin{bmatrix} m \\ m \end{bmatrix}, \begin{bmatrix} \Sigma_{tot} & \Sigma_{ac} \\ \Sigma_{ac} & \Sigma_{tot} \end{bmatrix} \right) - \\ &- \ln \mathcal{N}\left(\begin{bmatrix} w_i \\ w_j \end{bmatrix} | \begin{bmatrix} m \\ m \end{bmatrix}, \begin{bmatrix} \Sigma_{tot} & 0 \\ 0 & \Sigma_{tot} \end{bmatrix} \right) = \\ =& (w_i - m)^T Q (w_i - m) + (w_j - m)^T Q (w_j - m) + \\ &+ 2(w_i - m)^T P (w_j - m) + C \ , \end{aligned} \tag{28}$$

$$\Sigma_{tot} = VV^T + \Lambda^{-1} , \quad (29)$$
$$\Sigma_{ac} = VV^T , \quad (30)$$
$$H = (\Sigma_{tot} - \Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac}) , \quad (31)$$
$$Q = \Sigma_{tot}^{-1} - H^{-1} , \quad (32)$$
$$P = \Sigma_{tot}^{-1}\Sigma_{ac}H^{-1} , \quad (33)$$

and λ is a calibration value.

2.3.2 Estimation of the number of speakers

Let us assume that there are K_{max} ways of clustering a set W: $\{\Theta_1, \ldots, \Theta_{K_{max}}\}$, where K_{max} is the maximum possible number of speakers. The estimation of the true number of speakers can be obtained by solving (7). The main problem is to determine the likelihood function $P(W|\Theta)$ for any clustering Θ. The author proposes two different likelihood functions based on the developed PLDA model.

Likelihood function $\widetilde{L}_1$

The first likelihood function $\widetilde{L}_1$ is based on the "naive" assumption that the whole set of i-vectors W can be split into statistically independent pairs. Then a new set $W^* = \left\{ \begin{bmatrix} w_i \\ w_j \end{bmatrix} \right\}_{i,j=1,\, j>i}^{R}$ is considered. In this way the following likelihood function can be obtained:

$$\widetilde{L}_1(W^*|\Theta, V, \Lambda) = \sum_{i=1}^{R}\sum_{j=i+1}^{R} \ln P\left(\begin{bmatrix} w_i \\ w_j \end{bmatrix} |\Theta, V, \Lambda \right) =$$
$$= \sum_{i=1}^{R}\sum_{j=i+1}^{R} \begin{cases} -\frac{1}{2}\left(w_i^T H^{-1} w_i - 2w_i^T P w_j + w_j^T H^{-1} w_j\right) + C + \beta & , z_i = z_j \\ -\frac{1}{2}\left(w_i^T \Sigma_{tot}^{-1} w_i + w_j^T \Sigma_{tot}^{-1} w_j\right) + C & , z_i \neq z_j \end{cases} , \quad (34)$$

where z_i, z_j are the indicator vectors; $\beta = \ln P(H_s) - \ln P(H_d)$ is a calibration parameter. The increase of the β leads to preferring the H_s hypothesis and decreases the estimated number of speakers.

Likelihood function $\widetilde{L}_2$

The second likelihood function $\widetilde{L}_2$ is based on the classic PLDA assumption that i-vectors from different speakers are statistically independent. Ac-

cording to this assumption the likelihood function can be written as follows:

$$\begin{aligned}\tilde{L}_2(W|\Theta, V, \Lambda) =& \sum_{k=1}^{K} \ln P(w_{i_1^k}, w_{i_2^k}, \ldots, w_{i_{R_k}^k}|V, \Lambda) = \\ =& \frac{1}{2} \sum_{k=1}^{K} \left(\ln |\hat{\Sigma}_{y_k}| + \hat{y_k}^T \hat{\Sigma}_{y_k}^{-1} \hat{y_k} \right) + C \ ,\end{aligned} \tag{35}$$

where $I^k = \{i_1^k, \ldots, i_{R_k}^k\}$ is the set of indexes which belong to the speaker k; $\hat{\Sigma}_{y_k}$, $\hat{y_k}$ is the covariance matrix and the mean value of the a posteriori distribution of the speaker factors y_k:

$$\hat{\Sigma}_{y_k} = (R_k V^T \Lambda V + \alpha I)^{-1} \ , \tag{36}$$

$$\hat{y_k} = \hat{\Sigma}_{y_k} V^T \Lambda \sum_{r \in I^k} (w_r - m) \ , \tag{37}$$

α is the additional calibration parameter of the a priori distribution y_k, so that:

$$P(y_k) = \mathcal{N}(y_k|0, (\alpha I)^{-1}) \ . \tag{38}$$

The increase of α leads to a wider prior distribution and decreases the estimated number of speakers.

Additional parameters α and β

The additional parameters α and β are the key insight for applying the PLDA model for speaker number estimation. These parameters can be considered as an analogue of a decision threshold in a hierarchical clustering algorithm and provide a simple way of calibrating a diarization system under different test conditions. The impact of the parameters on the diarization results will be described in section 3.4.

3 Results

3.1 Evaluation protocol

3.1.1 Diarization scoring

The first metric for diarization scoring used in this thesis is the Diarization Error Rate (DER) proposed by the National Institute of Standards and

Technology (NIST) in "Rich Transcription Evaluation Project" [11], which is defined as:

$$DER = \frac{\sum_{seg} \left(T(seg) \cdot \max(N_{ref}(seg), N_{sys}(seg)) - N_{correct}(seg) \right)}{\sum_{seg} T(seg) \cdot N_{ref}(seg)} , \quad (39)$$

where $T(seg)$ is the speech segment duration; $N_{ref}(seg)$ is the number of speakers present in the speech segment seg according to the reference segmentation; $N_{sys}(seg)$ is the number of speakers present in the speech segment seg according to the diarization results; $N_{correct}(seg)$ is the number of correctly identifined speakers in the speech segment seg.

The DER can be decomposed into speaker error (E_{spkr}), that is, speaker time attributed to the wrong speaker, missed speaker error (E_{miss}), and false alarm speaker error (E_{FA}):

$$E_{FA} = \frac{\sum_{N_{sys}(seg) > N_{ref}(seg)} T(seg) \cdot (N_{sys}(seg) - N_{ref}(seg))}{\sum_{seg} T(seg) \cdot N_{ref}(seg)} , \quad (40)$$

$$E_{miss} = \frac{\sum_{N_{ref}(seg) > N_{sys}(seg)} T(seg) \cdot (N_{ref}(seg) - N_{sys}(seg))}{\sum_{seg} T(seg) \cdot N_{ref}(seg)} , \quad (41)$$

$$E_{spkr} = \frac{\sum_{seg} \left(T(seg) \cdot \min(N_{ref}(seg), N_{sys}(seg)) - N_{correct}(seg) \right)}{\sum_{seg} T(seg) \cdot N_{ref}(seg)} . \quad (42)$$

Normalized speaker error ($\widetilde{E}_{spkr}$) is proposed in this thesis for diarization performance evaluation:

$$\widetilde{E}_{spkr} = \frac{E_{spkr}}{1 - E_{miss}} \quad (43)$$

Another scoring metric used for evaluation is the Average Cluster Purity

(ACP), Average Speaker Purity (ASP) and their geometric mean K:

$$
\begin{aligned}
ACP_c &= \frac{\sum_{s=1}^{S} n_{sc}^2}{(\sum_{s=1}^{S} n_{sc})^2} \,, \\
ACP &= \frac{1}{N} \sum_{c=1}^{M} ACP_c \sum_{s=1}^{S} n_{sc} \,, \qquad (44) \\
ASP_s &= \frac{\sum_{c=1}^{M} n_{sc}^2}{(\sum_{c=1}^{M} n_{sc})^2} \,, \\
ASP &= \frac{1}{N} \sum_{s=1}^{S} ASP_s \sum_{c=1}^{M} n_{sc} \,, \qquad (45) \\
N &= \sum_{s=1}^{S} \sum_{c=1}^{M} n_{sc} \,, \\
K &= \sqrt{ACP \cdot ASP} \,, \qquad (46)
\end{aligned}
$$

where S is the number of speakers according to the reference segmentation; M is the number of clusters according to the diarization results; n_{sc} is the number of frames of cluster c attributed to speaker s.

3.1.2 Training data

The training data consists of audio recordings provided by NIST for Speaker Recognition Evaluations (SRE) in 2004, 2005, 2006 and 2008 [12]. This training data consists of recordings from three different channels: phone, microphone and distance microphone (interview microphone). The characteristics of the training data are presented in the Tables 1, 2.

Table 1: Number of speakers in training data

Data set	Number of recordings		Number of speakers	
	Male	Female	Male	Female
NIST-2004	1890	2648	122	184
NIST-2005	3733	5004	218	307
NIST-2006	4082	5172	353	477
NIST-2008	6615	10898	492	844
Total	16320	23722	1185	1812

Table 2: Distribution of channels in training data

Data set	Channel		
	Phone	Microphone	Distance microphone
NIST-2004	4538	0	0
NIST-2005	6004	2733	0
NIST-2006	6574	2680	0
NIST-2008	12191	1464	3858
Total	29307	6877	3858

3.1.3 Evaluation data

The following evaluation data sets were used:

- **NIST2008-ENG**

  This data set was formed by summing two channels of English phone conversation recordings from NIST SRE 2008 evaluation data [13]. It is important that this data set was excluded from the **NIST-2008** training set.

- **NIST2008-FOR**

  This data set was formed by summing two channels of non-English phone conversation recordings from NIST SRE 2008 evaluation data [13].

- **AMI-CORPUS**

  This data set was formed from the AMI Meeting Corpus [14] using the *Headset mix* subset. This test set contains 55 summed recordings of spontaneous speech of four speakers.

- **STC-MICG, STC-MIC1m, STC-MIC2m**

  This test data set was recorded at Speech Technology Center Ltd. [15] and contains recordings of a meeting of 2 speakers in three audio channels: headset (**STC-MICG**), microphone at a distance of 1 meter (**STC-MIC1m**), microphone at a distance of 2 meters (**STC-MIC2m**).

- **NIST2008-MONO**

  This test set was formed using one channel of recordings **NIST2008-ENG** and contains one speaker only.

- **NIST2008-QUAD**

  This is an artificial test set formed by means of concatenating dialog recordings from **NIST2008-ENG** which include different speakers. The recordings in this test data set contain exactly four speakers.

A brief overview of parameters of the evaluation sets are presented in the Table 3

Table 3: Parameters of the evaluation data sets

Data set	Languages	Channel	Number of record-ings	Number of speak-ers	Average recording duration, sec
NIST2008-ENG	English	phone	100	2	5 min
NIST2008-FOR	Chinese, Korean, Japanese, Thai, Hindi, Vietnamese	phone	100	2	5 min
AMI-CORPUS	English	headset	55	4	33 min
STC-MICG	Russian	headset	76	2	6 min 40 sec
STC-MIC1m	Russian	distance micro-phone	76	2	6 min 40 sec
STC-MIC2m	Russian	distance micro-phone	76	2	6 min 40 sec
NIST2008-MONO	English	phone	100	1	5 min
NIST2008-QUAD	English	phone	100	4	10 min

3.2 Front-end preprocessing

3.2.1 Voice Activity Detection

A Voice Activity Detection (VAD) based on energy analysis was used. The frequency range from 300Hz to 1500Hz was used for signal energy estimation on each frame with a 10 ms step and a 25 ms window. The energy decision threshold was calculated within a 10 second collar of the current frame. Speech segments shorter than 0.1 sec were excluded.

3.2.2 Voice features

The impact of different voice features (MFCC, PLP, LPC) was examined. It was shown that one of most effective voice features for the speaker diarization task was "raw" MFCC concatenated with PLP without any normalization or postprocessing techniques.

3.2.3 i-vectors extractor

UBM and T-matrix were trained through EM iterations. The UBM consists of 512 components with a diagonal covariance matrix. The dimension 100 of the T-matrix was used.

3.2.4 PLDA model estimation

The speech frames from each training recording were split into short speech segments ranging from 1 to 20 seconds in accordance with (5). The i-vectors obtained on these segments were used for PLDA model estimation (6). The dimension 50 of the matrix V and the full covariance matrix Λ were used.

3.3 Clustering for known number of speakers

Three methods of clustering speech segments were used for diarization performance evaluation:

- **VBA-TV**

  This method was proposed in [2] and based on the Variational Bayesian Analysis and Total Variability approach.

- **VBA-PLDA**

This method was described in Section 2.3.1. It is based on Variational Bayesian Analysis of the distribution of i-vectors extracted from short speech segments of audio recordings.

- **KERNEL K-MEANS**

  This method was briefly described in Section 2.3.1 and was used for first "rough" clustering of speech segments and subsequent initialization of the method **VBA-PLDA**.

The results of speaker diarization for a known number of speakers are presented in Table 4.

Table 4: Speaker error of diarization for known number of speakers

Method	$\widetilde{E}_{spkr}$					
	NIST2008-ENG	**NIST2008-FOR**	**AMI-CORPUS**	**STC-MICG**	**STC-MIC1m**	**STC-MIC2m**
VBA-TV	7.59	7.22	**13.89**	7.61	8.54	10.86
KERNEL K-MEANS + VBA-PLDA	6.68	**6.58**	21.63	6.00	10.78	12.65
KERNEL K-MEANS + VBA-PLDA + VBA-TV	**6.07**	**6.60**	14.02	**5.53**	**8.39**	**10.08**

As we can see, the best performance can be obtained when a combination of the proposed method **VBA-PLDA** and baseline method **VBA-TV** is used. This combination is achieved by using the clustering result of the method **VBA-PLDA** for the subsequent initialization of the **VBA-TV**.

3.4 Clustering for an unknown number of speakers

First, the dependence between ASP (45), ACP (44), K (46) and the proposed additional parameters α and β (2.3.2) is presented in figures 2, 1. The ACP, ASP are obtained by applying the proposed method for speech segment clustering with a fixed number of speakers ranging from 1 to 10 and further selection of the number of speakers using likelihood functions $\widetilde{L}_1$ (34) and $\widetilde{L}_2$ (35).

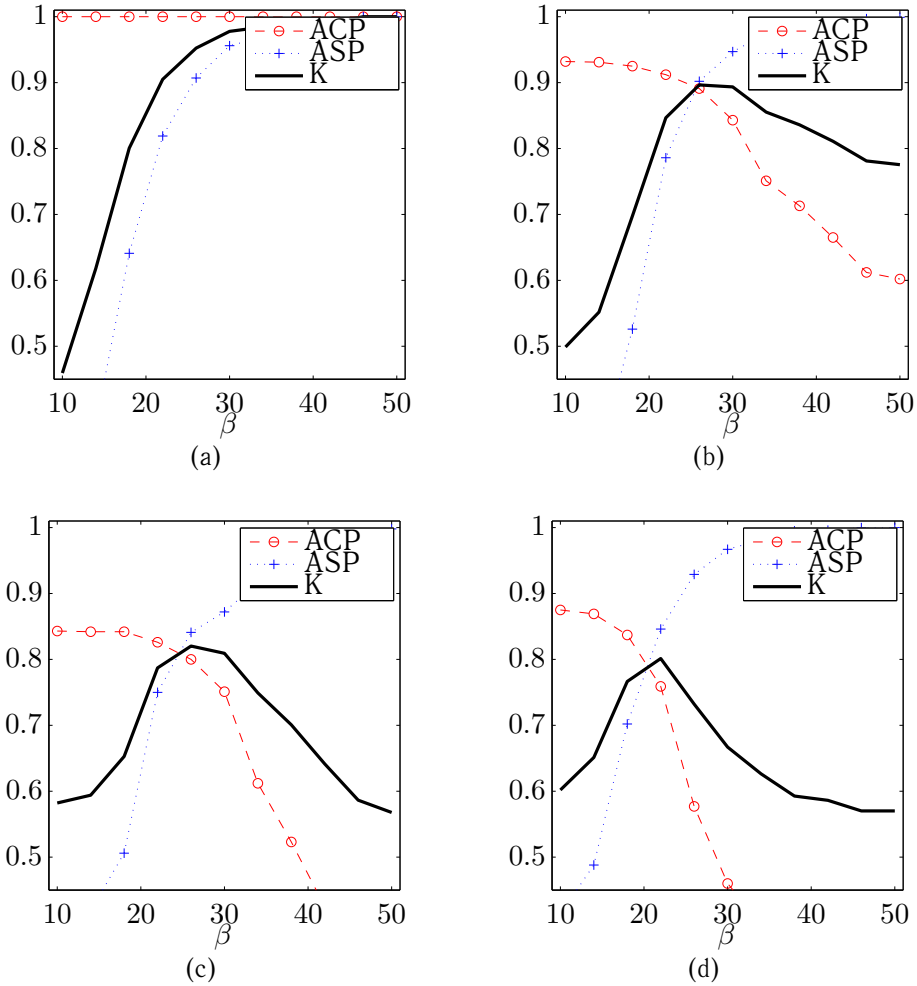

Figure 1: The diarization results (ASP, ACP, K) were obtained by using the likelihood function $\widetilde{L}_1$ for speaker number estimation with different values of the parameter β for the following test sets: (a) **NIST2008-MONO**; (b) **NIST2008-ENG**; (c) **NIST2008-QUAD**; (d) **AMI-CORPUS**

Next, the dependence between ASP and ACP for the proposed diarization methods when the number of speakers is unknown is presented on figure **??**. In addition, the values of ASP and ACP for the case when the number of speakers is known a priori are also presented. These curves are analogous to DET-curves and allow us to compare the effectiveness of proposed methods. As we can see, the ASP-ACP curve for the likelihood function $\widetilde{L}_2$ lies above the ASP-ACP curve for the likelihood function $\widetilde{L}_1$ and passes through the ASP-ACP point (black asterisk) obtained for

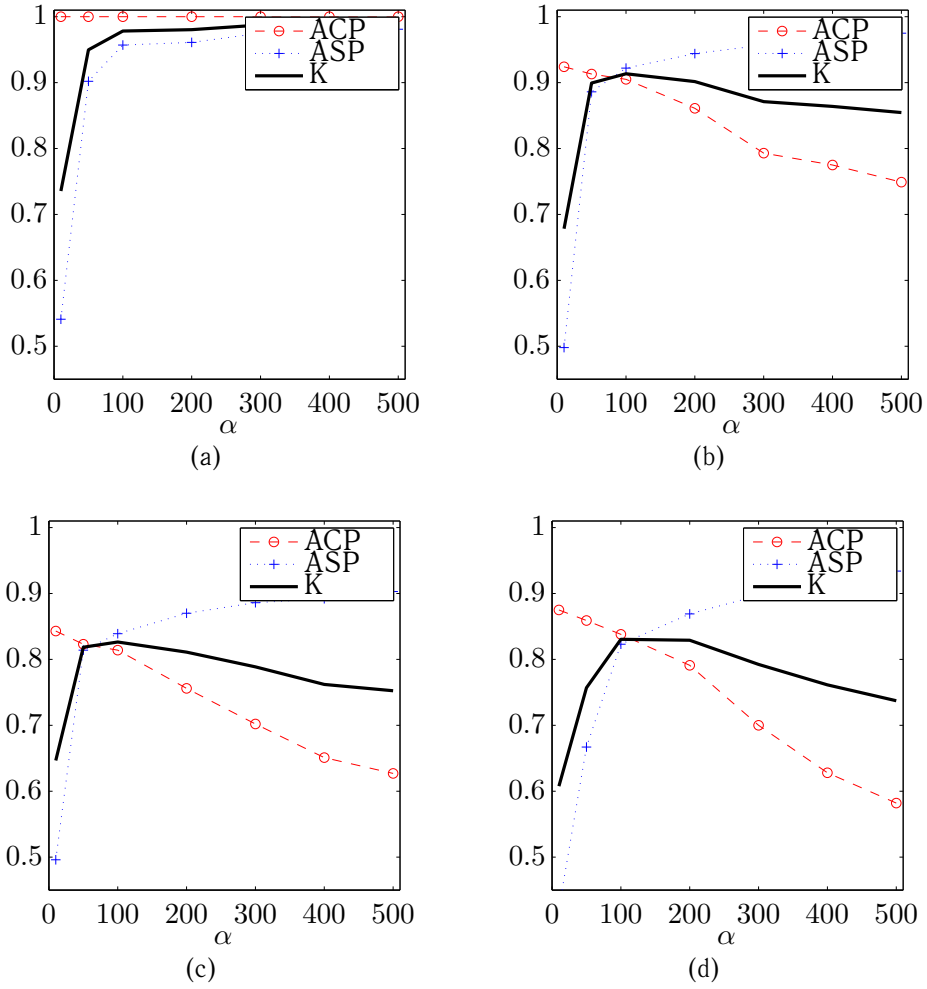

Figure 2: The diarization results (ASP, ACP, K) were obtained by using the likelihood function $\widetilde{L}_2$ for speaker number estimation with different values of the parameter α for the following test sets: (a) **NIST2008-MONO**; (b) **NIST2008-ENG**; (c) **NIST2008-QUAD**; (d) **AMI-CORPUS**

the "ideal" selection of the number of speakers. This demonstrates a high efficiency of speaker number estimation by using the proposed likelihood function $\widetilde{L}_2$.

Finally, the diarization results for an unknown number of speakers using $\widetilde{L}_2$ with the fixed parameter α are presented in Table 5.

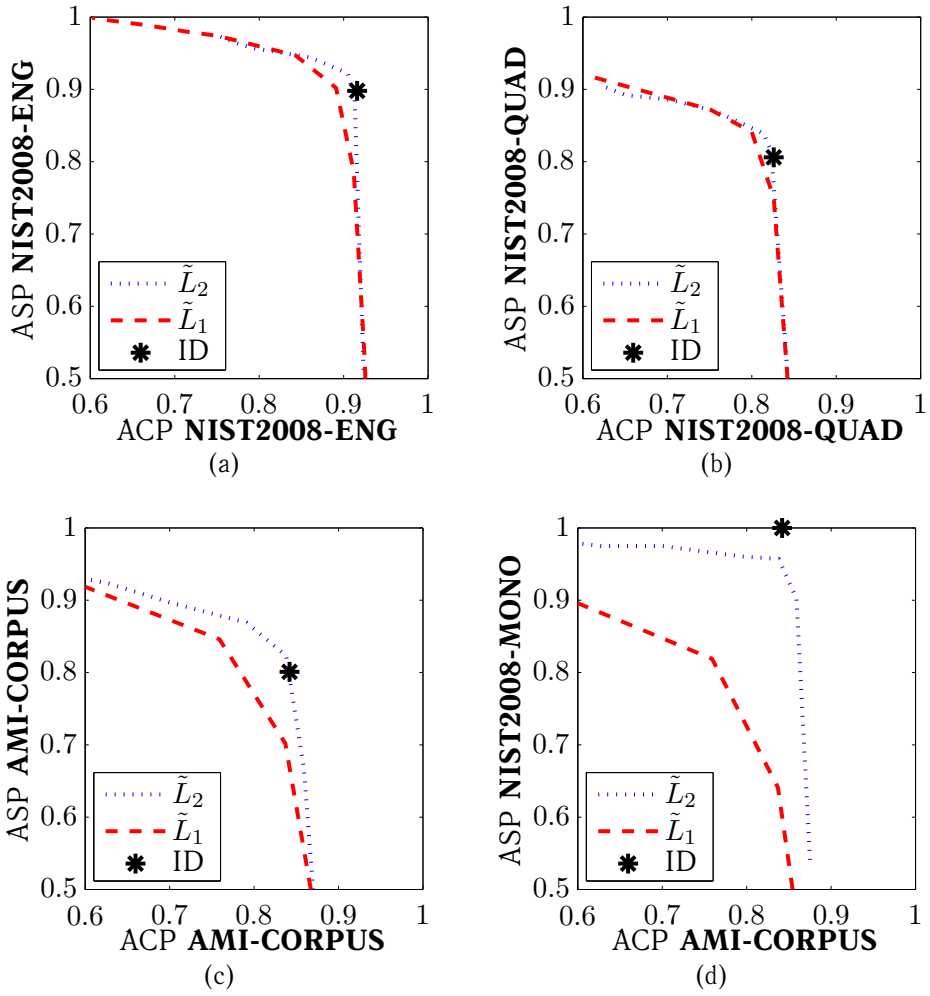

Figure 3: The dependence between ASP and ACP for the proposed methods of speaker number estimation and for the case when the number of speakers is known

4 Conclusion

This report presents a brief overview of a diarization system based on PDLA. The author proposes to divide the speaker diarization task into two subtasks: speaker diarization with a known number of speakers and estimation of the number of speakers. The report describes the original PLDA model and its application for these subtasks. The author introduces additional parameters of a priori distributions which provide a way to calibrate the developed diarization system. A series of experiments demonstrates the

Table 5: The diarization results for unknown number of speakers using $\widetilde{L}_2$ with fixed parameter α

		NIST2008-MONO	**NIST2008-ENG**	**NIST2008-FOR**	**AMI-CORPUS**	**STC-MICG**	**STC-MIC1m**	**STC-MIC2m**
α		100	100	100	100	100	100	100
Selected num. of speakers, %	1	93	10	15	0	40.8	25	49
	2	7	89	82	7	55.2	75	51
	3	0	1	3	22	4	0	0
	4	0	0	0	49	0	0	0
	5	0	0	0	22	0	0	0
	>5	0	0	0	0	0	0	0
ACP		1	0.905	0.867	0.838	0.840	0.811	0.762
ASP		0.957	0.922	0.911	0.823	0.946	0.894	0.937
K		0.978	0.913	0.889	0.830	0.891	0.852	0.845
$\widetilde{E}_{spkr}$		3.32	5.73	8.98	13.11	11.34	13.77	17.02

effectiveness of the proposed approach.

References

[1] O. Kudashev, *Система разделения дикторов на основе вероятностного линейного дискриминантного анализа*. PhD thesis, ITMO University, 2014. `https://isu.ifmo.ru/pls/apex/f?p=2005:0::DWNLD_F:NO::FILE:9916F6AC0AD4BDB0CA2DF830B0A90CEB`.

[2] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, pp. 1059–1070, Dec 2010.

[3] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," tech. rep., CRIM, 2008.

[4] S. Shum, N. Dehak, E. Chuangsuwanich, D. A. Reynolds, and J. R. Glass, "Exploiting intra-conversation variability for speaker diarization.," in *INTERSPEECH*, pp. 945–948, ISCA, 2011.

[5] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, Oct 2007.

[6] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey-2010*, (Brno, Czech Republic), 2010.

[7] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[8] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH*, pp. 249–252, 2011.

[9] P. Kenny, T. Stafylakis, P. Ouellet, M. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7649–7653, May 2013.

[10] I. S. Dhillon, Y. Guan, and B. Kulis, "Weighted graph cuts without eigenvectors: A multilevel approach," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 29, p. 2007, 2007.

[11] "NIST Rich Transcription Evaluation Project." http://www.itl.nist.gov/iad/mig/tests/rt/, 2009.

[12] "NIST Speaker Recognition Evaluation." http://nist.gov/itl/iad/mig/sre.cfm, 2012.

[13] "2008 NIST Speaker Recognition Evaluation." http://www.itl.nist.gov/iad/mig/tests/sre/2008/, 2008.

[14] "AMI Meeting Corpus." https://www.idiap.ch/dataset/ami/, 2014.

[15] "Speech Technology Center Ltd.." http://www.speechpro.com/.