PLDA-based Clustering for Speaker Diarization of Broadcast Streams

Jan Silovsky, Jan Prazak, Petr Cerva, Jindrich Zdansky, Jan Nouza

Institute of Information Technology and Electronics, Faculty of Mechatronics,
Technical University of Liberec, Czech Republic

{jan.silovsky,jan.prazak,petr.cerva,jindrich.zdansky,jan.nouza}@tul.cz

Abstract

This paper presents two approaches to speaker clustering based on Probabilistic Linear Discriminant Analysis (PLDA) in the speaker diarization task. We refer to the approaches as the multifold-PLDA approach and the onefold-PLDA approach. For both approaches, simple factor analysis model is employed to extract low-dimensional representation of a sequence of acoustic feature vectors – so called i-vectors – and these i-vectors are modeled using the PLDA model. Further, two-stage clustering with Bayesian Information Criterion (BIC) based approach applied in the first stage and the PLDA-based approach in the second stage is examined. We carried out our experiments using the COST278 multilingual broadcast news database. The best evaluated system yielded 42 % relative improvement of the speaker error rate over a baseline BIC-based system.

Index Terms: speaker diarization, PLDA, clustering, i-vectors

1. Introduction

Speaker diarization is the process of partitioning an input audio data into homogeneous segments according to a specific speaker identity (it solves the "who spoke when" task) and it is a useful preprocessing step in speech or speaker recognition and for indexing of audio archives. It can also improve the readability of automatic transcriptions. An inherent part of a speaker diarization system is a clustering module. The aim of clustering is to group segments of the same speaker together. In this paper, we investigate two clustering approaches based on the Probabilistic Linear Discriminant Analysis (PLDA).

The PLDA was initially introduced for the face recognition task [1] and it was recently successfully applied in the speaker detection task in the NIST 2010 Speaker Recognition Evaluation (SRE) [2]. Compared to the face recognition task based on comparison of two images of a defined resolution, speaker recognition operates with observations (sequences of feature vectors) of variable length and thus a projection to a fixed-dimensional feature vector must be performed. We apply a simple factor analysis model to extract low-dimensional representation of audio segments using so called i-vectors as proposed by [3].

Our motivation for utilization of PLDA stems from the following reasons. First, the PLDA model provides separation of speaker-specific and nuisance variability. Further, PLDA provides implicitly symmetric scoring. When deciding about whether two segments share the same identity or not, traditional speaker recognition methods usually employ a speaker model, trained using one of the segments, which is scored against the other segment. Cross score is computed for swapped segments [4] and symmetric score is then obtained as average of both scores. Finally, the PLDA model supports operation with multiple observations (speech segments in our case). In contrast, traditional speaker recognition methods usually handle multiple segments by merging them into one segment, or equivalently by summation of statistics derived for each segment.

2. Speaker diarization system

Our speaker diarization system consists of three basic modules. First, after feature vectors are extracted, speech activity detection (SAD) is applied. Then, speaker change points are detected by a speaker segmentation module. Finally, segments of the same speakers are clustered and speaker diarization is provided. All components of the system use classic Mel-frequency cepstral coefficient (MFCC) features.

The speech activity detector has two parts - an energy based detector with an adaptive threshold and a Gaussian Mixture Model (GMM) based detector. The aim of the former is to remove silent parts from the signal, while the latter does the same for other non-speech events (especially for music and noise).

The aim of the speaker segmentation module is to find speaker change points in previously identified speech segments. For that purpose, we use method based on the Bayesian Information Criterion (BIC) introduced in [5]. This technique searches for one change point within an adaptive (variable-length) window that moves subsequently through all the speech segments.

The clustering module uses bottom-up clustering (a.k.a. hierarchical agglomerative clustering) which is predominant approach for speaker clustering. First, a similarity measure between all pairs of speech segments is computed. Next, until the stopping criterion is met, the most similar pair of speech segments (clusters) is iteratively merged into a new cluster and the similarity measure between the new cluster and all remaining speech segments (clusters) is recomputed.

3. Clustering methods

3.1. BIC-based clustering

Probably the most popular clustering similarity measure is a metric based on the BIC [5]. The BIC-based criterion compares the BIC statistic of clusters g_1 and g_2 with the BIC statistic of the cluster g which is formed by merging of the g_1 and the g_2. We apply a local BIC measure which is defined as

$$\begin{aligned}\Delta BIC(g_1, g_2) = &(N_1 + N_2) log\,|\mathbf{\Sigma}| - N_1 log\,|\mathbf{\Sigma}_1| \\ &- N_2 log\,|\mathbf{\Sigma}_2| - \alpha P\end{aligned} \quad (1)$$

where N is the number of frames, $\mathbf{\Sigma}$ is the full covariance matrix of the data and P is the penalty

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right) log(N_1 + N_2) \quad (2)$$

where d is the dimension of feature vectors and α is a penalty weight.

In the clustering process, two clusters with the lowest ΔBIC value are merged together. If a minimal distance between any pair of clusters is higher than a certain threshold λ (typically zero), the stopping criterion is met.

3.2. I-vectors extraction

Before we can approach PLDA-based clustering, a fixed-dimensional representation of a segment of variable length must be extracted. We employ a simple factor analysis model as proposed by [3]. Let's assume a GMM trained on data pooled from many speakers. This model is typically referred to as the Universal Background Model (UBM). The term *supervector* is used to refer to a high-dimensional vector obtained by concatenation of mean vectors of components of a GMM. Let $\boldsymbol{s}$ be a supervector representing a speech segment. The speaker-and segment-specific supervector for j'th segment of a speaker s is defined using the generative model

$$\boldsymbol{s}_{s,j} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{x}_{s,j} \tag{3}$$

where $\boldsymbol{m}$ is the speaker-and segment-independent supervector (obtained from the UBM), the $\boldsymbol{T}$ is a rectangular matrix of low rank and $\boldsymbol{x}_{s,j}$ is a random vector having standard normal distribution $\mathcal{N}[0, \boldsymbol{I}]$. The matrix $\boldsymbol{T}$ defines a total variability space and components of the vector $\boldsymbol{x}$ are the total factor loadings. Following the terminology of [3] we refer to the vector $\boldsymbol{x}$ as the *i-vector*.

A projection from a sequence of feature vectors representing a speech segment to the i-vector space is provided by computation of a Maximum A Posterior (MAP) point estimate of the total factor loadings based on zero-and first-order sufficient statistics gathered employing the UBM [6]. Having a fixed-dimensional representation we can apply the PLDA. Motivated by [3] which deals with application of cosine distance scoring for speaker recognition using i-vectors, we apply unit length normalization of i-vectors.

3.3. Probabilistic linear discriminant analysis

Now we put aside the assumption of i-vectors having distribution $\mathcal{N}[0, \boldsymbol{I}]$ and consider another factor analysis model that aims to separate speaker-specific and nuisance variability in the i-vector space. The PLDA model defines generation process of the i-vector $\boldsymbol{x}_{s,j}$ as

$$\boldsymbol{x}_{s,j} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{z}_{s,j} + \boldsymbol{\epsilon}_{s,j} \tag{4}$$

where $\boldsymbol{\mu}$ is the overall speaker-and segment-independent mean of the vectors in the training dataset, columns of the matrix $\boldsymbol{V}$ define bases for the subspace where the speaker-specific variability resides (the columns are referred to as eigenvoices) and columns of the matrix $\boldsymbol{U}$ define bases for the nuisance variability subspace (the columns are referred to as eigenchannels[1]). The term $\boldsymbol{\epsilon}_{s,j}$ represents unexplained residual variability which is defined by the diagonal covariance matrix $\boldsymbol{\Sigma}$. The components of the vector $\boldsymbol{y}_s$ are the eigenvoice factor loadings and components of the vector $\boldsymbol{z}_{s,j}$ are the eigenchannel factor loadings. Both loadings vectors are assumed to have standard normal distribution, i.e. $p(\boldsymbol{y}_s) = \mathcal{N}_{\boldsymbol{y}}[0, \boldsymbol{I}]$ and $p(\boldsymbol{z}_{s,j}) = \mathcal{N}_{\boldsymbol{z}}[0, \boldsymbol{I}]$. Please note that the term $\boldsymbol{V}\boldsymbol{y}_s$ depends only on the identity of the speaker and not on the particular segment.

The model represented by Eq. (4) can be expressed in terms of conditional probability as follows:

$$p(\boldsymbol{x}_{s,j}|\boldsymbol{y}_s, \boldsymbol{z}_{s,j}, \theta) = \mathcal{N}_{\boldsymbol{x}}[\boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{z}_{s,j}, \boldsymbol{\Sigma}] \tag{5}$$

where θ represents the set of parameters $\{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{\Sigma}\}$ that are estimated using the Expectation Maximization (EM) algorithm on the background data. These parameters remain fixed during the recognition phase.

In recognition, we aim to compute the likelihood of the observed data. Considering the clustering task, evaluation of the likelihood $p(\boldsymbol{x}_{1\ldots N})$ that N observations $\boldsymbol{x}_{1\ldots N}$ share the same identity is particularly of our interest. Combining the PLDA generative models for all N observations sharing the identity $\boldsymbol{y}$ we get a compound model:

$$\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\\ \vdots\\ \boldsymbol{x}_N\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\\ \vdots\\ \boldsymbol{\mu}\end{bmatrix} + \begin{bmatrix}\boldsymbol{V} & \boldsymbol{U} & 0 & \ldots & 0\\ \boldsymbol{V} & 0 & \boldsymbol{U} & \ldots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{V} & 0 & 0 & \ldots & \boldsymbol{U}\end{bmatrix} \begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{z}_1\\ \boldsymbol{z}_2\\ \vdots\\ \boldsymbol{z}_N\end{bmatrix} + \begin{bmatrix}\boldsymbol{\epsilon}_1\\ \boldsymbol{\epsilon}_2\\ \vdots\\ \boldsymbol{\epsilon}_N\end{bmatrix} \tag{6}$$

which we can rewrite as:

$$\boldsymbol{x}' = \boldsymbol{\mu}' + \boldsymbol{A}\boldsymbol{w} + \boldsymbol{\epsilon}' \tag{7}$$

because $p(\boldsymbol{w}) = \mathcal{N}_{\boldsymbol{w}}[0, \boldsymbol{I}]$, the likelihood of the compound model is given as:

$$p(\boldsymbol{x}_{1\ldots N}) = p(\boldsymbol{x}') = \mathcal{N}_{\boldsymbol{x}'}[\boldsymbol{\mu}', \boldsymbol{A}\boldsymbol{A}^T + \boldsymbol{\Sigma}'] \tag{8}$$

where $\boldsymbol{\Sigma}'$ is block diagonal matrix whose diagonal blocks are $\boldsymbol{\Sigma}$. Eq. 8 thus represents the likelihood that the segments represented by i-vectors $\{\boldsymbol{x}_1, \ldots \boldsymbol{x}_N\}$ all share the same identity. Please note that no point estimates of hidden variables $\boldsymbol{y}$ or $\{\boldsymbol{z}_1, \ldots \boldsymbol{z}_N\}$ are used for the likelihood computation, instead the hidden variables are integrated out [1].

3.4. Multifold-PLDA approach

In the multifold-PLDA approach, a cluster is represented by a set of i-vectors corresponding to the segments assigned to the cluster. Let $\boldsymbol{X}^{(g)} = \{\boldsymbol{x}^{(g)}_{1\ldots J^{(g)}}\}$ be the set of $J^{(g)}$ i-vectors representing a cluster g and $\boldsymbol{x}'^{(g)}$ a compound vector formed by concatenation of the i-vectors. In the clustering process, we aim to compare the likelihood of two competing models. Under the first model $\mathcal{M}_0$, clusters belong to different speakers and thus they have different speaker factor loadings $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. While under the second model $\mathcal{M}_1$, two clusters belong to the same speaker and thus have the same speaker factor loadings $\boldsymbol{y}$. The criterion used to decide whether the data are more likely represented by the model with a shared identity or by the model with different identities is based on the log-likelihood ratio:

$$LLR = \log \frac{p(\boldsymbol{x}'^{(1)}, \boldsymbol{x}'^{(2)}|\mathcal{M}_1)}{p(\boldsymbol{x}'^{(1)}, \boldsymbol{x}'^{(2)}|\mathcal{M}_0)} \tag{9}$$

Because the variables $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are independent under the model $\mathcal{M}_0$, the likelihood can be broken down into

$$p(\boldsymbol{x}'^{(1)}, \boldsymbol{x}'^{(2)}|\mathcal{M}_0) = p(\boldsymbol{x}'^{(1)}|\mathcal{M}_0)p(\boldsymbol{x}'^{(2)}|\mathcal{M}_0) \tag{10}$$

The likelihood for the model $\mathcal{M}_1$ is given as follows:

$$p(\boldsymbol{x}'^{(1)}, \boldsymbol{x}'^{(2)}|\mathcal{M}_1) = p(\boldsymbol{x}'|\mathcal{M}_1) \tag{11}$$

[1] We adopt the terminology used in speaker recognition where channel variability is usually supposed to represent not only variance between telephone and microphone speech but all the nuisance variability.

where $\boldsymbol{x}'$ is formed by concatenation of vectors $\boldsymbol{x}'^{(1)}$ and $\boldsymbol{x}'^{(2)}$. Likelihoods on the right-hand side of Eqs. (10) and (11) correspond to models with a single identity and are computed according to (8).

In the clustering process, the two clusters with the highest LLR value are merged together. If a maximum LLR value for any pair of clusters is lower than a certain threshold λ, estimated on the development data, the stopping criterion is met.

3.5. Onefold-PLDA approach

In the onefold-PLDA approach, a cluster is represented by a single i-vector. Sufficient statistics gathered employing the UBM for each segment assigned to the cluster are summed together and a MAP point estimate of the total factor loadings extracted based on these summed statistics. Although, compared to the multifold-PLDA system, an i-vector must be extracted every time a new cluster is formed, the onefold-PLDA system is less computational expensive as only one i-vector per a cluster participates in the likelihood computation.

Likewise in the multifold approach, the clustering process is driven by the LLR values between clusters.

4. Experiments and results

4.1. Datasets

Experiments were carried out using the COST278 multilingual pan-European broadcast news database [7]. The database comprises broadcast news recordings in 9 languages. Authors of the database have divided the data for each language into a training set (containing about two hours) and a test set (containing about one hour).

We divided the data into three datasets. The first set contained all COST278 Croatian, Czech, Hungarian, Portuguese and Slovak training data giving in total 11.5 hours of audio. This set was used for training of the UBM and estimation of the total variability space and parameters of the PLDA model. The second set, consisting of 13 shows of various lengths (in the range from 8.5 to 53.8 minutes) drawn also from the COST278 training data giving in total 5.89 hours, was used as the development set for tuning of system parameters. Particularly for estimation of segmentation and clustering stopping thresholds. Finally, the third set was used as the test set in our experiments. The set consisted of 15 shows of various lengths (in the range from 4.1 to 53.2 minutes) drawn from the COST278 test data giving in total 6.34 hours. The development and test data were limited to 5 languages: Belgian Dutch, Czech, Hungarian, Slovenian and Slovak. The streams in COST278 corpus contain also commercials which are not annotated. The commercials were thus removed from the streams used in development and test sets.

4.2. Evaluation metrics

Performance of diarization systems is usually evaluated by the Diarization Error Rate (DER) as the primary metric. The DER was defined by the National Institute of Standards and Technology (NIST) [8] and it can be decomposed as:

$$DER = SPKE + FA + MISS \qquad (12)$$

where the SPKE represents the speaker error rate, the FA is the speech false alarm error rate and the MISS is the missed speech error rate. The SPKE reflects the amount of speech data that is attributed to a wrong speaker given the optimum speaker mapping between a system output and a reference diarization. The FA reflects the amount of non-speech segments that were recognized as speech and the MISS reflects the amount of speech segments that were recognized as non-speech. Because all our evaluated systems share the same SAD and speaker segmentation modules, we use the SPKE as the primary metric. The NIST scoring tool[2] was employed to compute the metrics for our experiments. Likewise in [8], a forgiveness collar of 0.25 s (both + and -) was not scored around each boundary.

4.3. Baseline system

The SAD achieved FA of 0.8 % and MISS of 3.2 %. We found that higher value of the MISS is caused by inaccuracy of reference annotations. The average length of speech segments after segmentation was 3.6 s.

The baseline system employs the BIC-based clustering approach. First, performance for different values of the BIC penalty weight α was evaluated. We found that the systems using a value of the stopping threshold λ estimated on the development data yielded better performance than the systems operating with zero value of the threshold. The best performance was provided by the system using penalty weight α of 4.0 and the stopping threshold λ of 1268.8. The system achieved SPKE of 24.9 % which corresponds to the DER of 28.9 %. These results are considered as baseline.

4.4. PLDA system training data

The UBM with 1024 components was trained using data from 1007 speakers (2530 segments, 11.5 hours). The total variability space was estimated using a subset of the UBM training data resulting from the condition of minimal length of a segment of 3 seconds and using at most eight segments per speaker. This resulted in 2050 segments (10.2 hours) from 909 speakers. The eigenvoices and eigenchannels were jointly estimated [1] using data from speakers for which at least three segments of minimal length of 3 seconds are available, in total 1528 segments (7.5 hours) from 280 speakers were used. The average length of segments used in training is 17.8 s. For training of all subspaces, we employed the EM-algorithm proposed by [9] which performs both maximum likelihood and minimum divergence update at each iteration.

4.5. Multifold-PLDA system

Various configurations were examined differing in the number of Gaussians in the UBM, dimension of the total variability space and the number of eigenvoices and eigenchannels in the PLDA model. Table 1 shows results for two best performing configurations. The UBM with 256 Gaussians was used to extract the sufficient statistics in both cases. Although non-symmetric numbers of eigenvoices and eigenchannels were also examined, symmetric configurations always yielded better performance. The system employing 400-dimensional i-vectors and the PLDA model with 200 eigenvoices and 200 eigenchannels yielded 36 % reduction of the SPKE.

Table 1: *Results for the multifold-PLDA system.*

rk($\boldsymbol{T}$)	rk($\boldsymbol{V}$)	rk($\boldsymbol{U}$)	SPKE [%]	rel. impr. [%]
300	150	150	17.2	30.9
400	200	200	15.9	36.1

[2] http://itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl

4.6. Onefold-PLDA system

Table 2 summarizes results for the best performing setups of the onefold-PLDA system (again the UBM with 256 Gaussians was employed) and shows that the system also outperforms the baseline system. However, the performance improvement is of much smaller extent compared to the multifold-PLDA system. We attribute this to the loss of information caused by summation of the sufficient statistics. In case that the clusters belonging to the same speaker are merged together, we obtain a better estimate of i-vector components by virtue of summation of the statistics over the clusters since the summation averages out the intra-speaker variability. In contrast, when two clusters belonging to different speakers are merged, we obtain an i-vector belonging to a synthesized identity. This seems to have more impact than a contamination of a set of i-vectors representing a cluster by i-vector belonging to a different speaker which would occur in case of the multifold-PLDA system.

Table 2: *Results for the onefold-PLDA system.*

rk($\boldsymbol{T}$)	rk($\boldsymbol{V}$)	rk($\boldsymbol{U}$)	SPKE [%]	rel. impr. [%]
300	100	100	22.5	9.6
300	150	150	21.8	12.4
400	200	200	23.0	7.6

4.7. Two-stage clustering

We hypothesize that the MAP point estimate of the total factor loadings (i-vectors) for segments of short duration cannot be estimated reliably which may harm the clustering process particularly at early phases. This problem relates at various extent to both PLDA-based systems. To cope with the problem we employ two-stage clustering. In the first stage, we use BIC-based clustering with zero value of the stopping threshold λ and value of the BIC penalty weight α set so as to under-cluster the segments. In the next stage the PLDA-based clustering is applied. Table 3 shows achieved results.

Table 3: *Results for two-stage clustering.*

Multifold-PLDA system			Onefold-PLDA system		
BIC α	SPKE [%]	rel. impr. [%]	BIC α	SPKE [%]	rel. impr. [%]
rk($\boldsymbol{T}$) = 300, rk($\boldsymbol{V}$) = 150, rk($\boldsymbol{U}$) = 150					
2.0	17.0	31.7	2.0	19.2	22.9
2.5	18.3	26.5	2.5	16.0	35.7
3.0	18.9	24.1	3.0	14.5	41.8
rk($\boldsymbol{T}$) = 400, rk($\boldsymbol{V}$) = 200, rk($\boldsymbol{U}$) = 200					
2.0	14.9	40.2	2.0	19.0	23.7
2.5	14.7	41.0	2.5	16.6	33.3
3.0	15.1	39.4	3.0	16.9	32.1

Significant effect of the two-stage clustering scenario was observed particularly for the onefold-PLDA system. All evaluated setups of the system provided performance improvement. The system employing 300-dimensional i-vectors and the PLDA model with 150 eigenvoices and 150 eigenchannels achieved the best overall SPKE of 14.5 % (42 % relative reduction) for the BIC penalty weight of 3.0 used at the first stage.

For the multifold-PLDA system, rather minor effect of the two-stage clustering was observed for both system's configurations from Table 1. Better performance was yielded by the larger system. For the BIC penalty weight of 2.5, the system achieved SPKE of 14.7 % (41 % relative reduction). Compared to the onefold-PLDA system, the multifold-PLDA seems to provide better performance for more under-clustered segments which corresponds to the performance of both systems in the one-stage clustering scenario. However, two-stage clustering scenario improves performance for both systems.

5. Conclusions

In this paper we have described our speaker diarization system and presented two speaker clustering approaches based on the PLDA. The system using the first approach, denoted as the multifold-PLDA system, outperformed the baseline system based on the BIC relatively by 36 % in terms of speaker error rate. The onefold-PLDA system based on the second presented approach yielded 12 % performance improvement over the baseline. We argue that the i-vectors representing the speech segments cannot be estimated reliably for short segments and employ two-stage clustering. The first stage uses BIC-based clustering to under-cluster the segments and PLDA-based clustering is performed at the second stage. Significant effect of the two-stage clustering was observed particularly for the onefold-PLDA system. The onefold-PLDA system used in two-stage clustering scenario achieved the best overall speaker error rate of 14.5 % which corresponds to 42 % relative improvement over the baseline error rate of 24.9 %.

6. Acknowledgements

The research described in this paper was supported by the Technology Agency of the Czech Republic (project no. TA01011204) and by the Czech Science Foundation - GACR (project no. P103/11/P499).

7. References

[1] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings ICCV 2007*, Rio de Janeiro, Brazil, October 2007.

[2] N. Brummer, L. Burget, P. Kenny, P. Matějka, E. V. de, M. Karafiát, M. Kockmann, O. Glembek, O. Plchot, D. Baum, and M. Senoussauoi, "ABC system description for NIST SRE 2010," in *Proc. NIST 2010 Speaker Recognition Evaluation*. Brno University of Technology, 2010, pp. 1–20.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788 –798, May 2011.

[4] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and bic for speaker diarization," in *Interspeech'05, ISCA*, Lisbon, September 2005.

[5] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings 1998 DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 127–132.

[6] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Processing*, vol. 13, May 2005.

[7] A. Vandecatseye *et al.*, "The COST278 pan-European broadcast news database," 2004, pp. 873–876.

[8] NIST, "The 2009 (RT-09) rich transcription meeting recognition evaluation plan," 2009.

[9] N. Brummer, "The EM algorithm and minimum divergence," October 2009, unpublished. [Online]. Available: http://niko.brummer.googlepages.com/EMandMINDIV.pdf