INFORMATION THEORETIC CLUSTERING FOR UNSUPERVISED DOMAIN-ADAPTATION

Subhadeep Dey[1,2]*, Srikanth Madikeri*[1] *and Petr Motlicek*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{subhadeep.dey, srikanth.madikeri, petr.motlicek}@idiap.ch

ABSTRACT

The aim of the domain-adaptation task for speaker verification is to exploit unlabelled target domain data by using the labelled source domain data effectively. The i-vector based Probabilistic Linear Discriminant Analysis (PLDA) framework approaches this task by clustering the target domain data and using each cluster as a unique speaker to estimate PLDA model parameters. These parameters are then combined with the PLDA parameters from the source domain. Typically, agglomerative clustering with cosine distance measure is used. In tasks such as speaker diarization that also require unsupervised clustering of speakers, information-theoretic clustering measures have been shown to be effective. In this paper, we employ the Information Bottleneck (IB) clustering technique to find speaker clusters in the target domain data. This is achieved by optimizing the IB criterion that minimizes the information loss during the clustering process. The greedy optimization of the IB criterion involves agglomerative clustering using the Jensen-Shannon divergence as the distance metric. Our experiments in the domain-adaptation task indicate that the proposed system outperforms the baseline by about 14% relative in terms of equal error rate.

Index Terms— Speaker verification, Domain adaptation, Information theoretic measures, PLDA model.

1. INTRODUCTION

The i-vector based PLDA system requires large collection of labelled data (speaker labels) to deliver state-of-the-art performance [1]. However in realistic applications, it might be too expensive to provide labelled speaker data for every domain of interest. For instance, we consider the task of speaker verification on recordings from social media (e.g. Youtube, Facebook). In such a problem, a large amount of unlabelled data from the target (e.g. Youtube) domain may be provided while the system is built using telephone corpora (e.g. speaker-labelled recordings obtained from the Switchboard (SWB) database [2]). The workshop on domain-adaptation[1] for speaker verification targeted this problem by creating two development datasets, namely labelled data from Switchboard database and unlabelled NIST Speaker Recognition Evaluation (SRE) data from SRE04 to SRE08.

In the literature, many approaches have been explored to adapt the speaker recognition system to target domain data. Most of these approaches focus on reducing the mismatch at the model level. Particularly, in the i-vector PLDA based speaker verifications systems, adapting the PLDA parameters has been shown to be successful [3, 4, 5]. Another recent work at the model level uses a collection of whitening matrices for (length) normalizing the i-vectors [6]. However, the whiteners are estimated from the source domain. Alternatively, the inter-dataset variation is compensated at the i-vector level using techniques such as Nuisance Attribute Projection (NAP) [7].

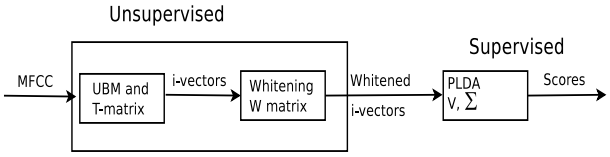

Fig. 1. Traditional framework of the speaker verification system.

Among all the above mentioned approaches, the target domain data is best used when adapting PLDA model parameters. To adapt these parameters, the target domain data is clustered and each cluster is assumed to represent a unique speaker. These speaker labels are then used to adapt the source domain PLDA parameters to the target domain [5, 6].

Speaker clustering on the unlabelled data is clearly a critical process that determines the success of model adaptation. A similar task where this problem occurs is speaker diarization, which labels the speaker segments in a speech recording in an unsupervised fashion. Commonly used model-based techniques use the Hidden Markov Model/Gaussian Mixture Model (HMM/GMM) [8, 9] to model the entire conversation. Alternatively, a non-parametric method using the Information Bottleneck (IB) method has been used [10]. Unlike HMM/GMM that clusters GMMs estimated from long segments of the audio, the IB method divides input speech into relatively short (~ 2.5 s) segments which are then iteratively merged using IB criterion to obtain the final speaker segments. The IB criterion minimizes the information loss in the clustering process while simultaneously finding compact representation of the data. In this paper, we explore IB method to perform speaker clustering on the unlabelled target domain data to be subsequently applied in speaker verification. In this work, we apply IB to cluster i-vectors to obtain speaker labels from the unlabelled (target domain) data. The distance metric for IB clustering uses Jensen-Shannon (JS) divergence which is shown to outperform conventional metrics.

The paper is organized as follows: Sections 2 and 3 describe the framework for the domain-adaptation task and the baseline speaker verification system, respectively. Section 4 presents the proposed information theoretic approach and Section 5 describes the experimental section. Finally, the paper is concluded in Section 6.

[1]http://www.clsp.jhu.edu/workshops/archive/ws13-summer-workshop/groups/spk-13/

2. DOMAIN-ADAPTATION FRAMEWORK

As shown in Figure 1, the training process of the traditional (i.e., i-vector PLDA) speaker verification module can be divided into two phases:

- **Unsupervised Phase**: In this phase the parameters of the Universal Background Model (UBM), Total variability matrix ($\mathbf{T}$) and the Whitening matrix ($\mathbf{W}$) are estimated. With reference to domain-adaptation, it was observed in [3] that using the target domain data to estimate parameters of the UBM and T-matrix does not improve the performance of the speaker verification system, however computing the whitening matrix on the target domain data significantly improves the performance.
- **Supervised Phase**: It consists of training the PLDA model parameters which requires speaker labels and multiple occurrences of the speaker.

3. BASELINE SPEAKER VERIFICATION SYSTEM

In simplified PLDA model [11], an i-vector ($\mathbf{x}$) is decomposed into speaker factor and residual as follows:

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{V}\mathbf{y} + \epsilon, \tag{1}$$

where the matrix $\mathbf{V}$ represents inter-class variability, $\mathbf{y}$ is the latent variable which follows a Gaussian distribution with zero mean ($\mathbf{0}$) and identity covariance matrix ($\mathbf{E}$); the residual ϵ follows the Gaussian distribution with zero mean ($\mathbf{0}$) and full covariance matrix ($\boldsymbol{\Sigma}$), $\boldsymbol{\mu}$ is the mean of the i-vectors. The distance between two i-vectors ($\mathbf{x}_1$ and $\mathbf{x}_2$) is computed as:

$$S(\mathbf{x}_1, \mathbf{x}_2) = \frac{p(\mathbf{x}_1, \mathbf{x}_2|H_s)}{p(\mathbf{x}_1, \mathbf{x}_2|H_d)}, \tag{2}$$

where the hypothesis H_s is that the two i-vectors share the same speaker latent variable ($\mathbf{y}$) and the hypothesis H_d is that the two i-vectors do not share the same latent variable. A closed form solution of Equation 2 can be found in [11]. In the PLDA model, an i-vector follows a Gaussian distribution as $\mathcal{N}(\mu, \boldsymbol{\Phi} + \boldsymbol{\Sigma})$, where $\boldsymbol{\Phi} = \mathbf{V}\,\mathbf{V}^t$. The adaptation procedure, as described in [12], of the source ($\mathbf{V}_{\text{out}}$ and $\boldsymbol{\Sigma}_{\text{out}}$) and target ($\mathbf{V}_{\text{in}}$ and $\boldsymbol{\Sigma}_{\text{in}}$) domain PLDA model are given by the following equations:

$$\begin{aligned} \boldsymbol{\Sigma} &= \alpha\boldsymbol{\Sigma}_{\text{out}} + (1-\alpha)\boldsymbol{\Sigma}_{\text{in}}, \\ \boldsymbol{\Phi} &= \alpha(\mathbf{V}_{\text{out}}\mathbf{V}'_{\text{out}}) + (1-\alpha)(\mathbf{V}_{\text{in}}\mathbf{V}'_{\text{in}}), \end{aligned} \tag{3}$$

where ($\boldsymbol{\Sigma}$, $\boldsymbol{\Phi}$) are adapted parameters of the new PLDA model and the parameter α ($\in$ [0,1]) balances the contribution of the source and target domain PLDA model [4].

To estimate the PLDA model parameters $\mathbf{V}_{\text{in}}$ and $\boldsymbol{\Sigma}_{\text{in}}$ on the target domain data, speaker labels are required. Thus, the data is clustered and each cluster is assigned to a unique speaker label. For instance, agglomerative clustering of i-vectors using a simple cosine distance metric is shown to provide good speaker labels for successful PLDA adaptation [4] and is used as the baseline system.

4. INFORMATION BOTTLENECK METHOD FOR SPEAKER CLUSTERING

In the IB framework, the input variable $\mathbf{U}$ is associated with a relevance variable ($\mathbf{R}$), which signifies some information that needs to be preserved during clustering. For instance in speech, this information can be related to acoustic classes, speaker identity, etc. The IB finds K clusters $\mathbf{C} = \{C_1, C_2, \cdots, C_{\text{K}}\}$ of the input data $\mathbf{U}$ such that (i) it is the most compact representation of the data, and (ii) it preserves most of the information in the relevance variables after clustering. Mathematically, this is equivalent to maximizing the following objective function ($\mathcal{F}$):

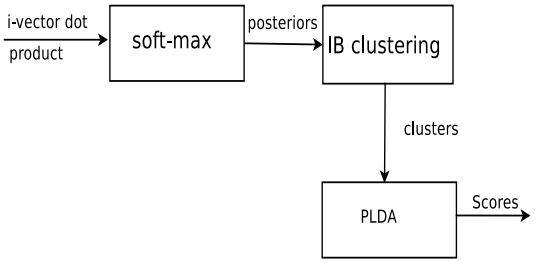

Fig. 2. Framework of the proposed system.

$$\mathcal{F} = \text{I}\left(\mathbf{R}, \mathbf{C}\right) - \frac{1}{\beta}\text{I}\left(\mathbf{U}, \mathbf{C}\right), \tag{4}$$

where I is the Mutual Information (MI) between two random variables. The first term in the right hand side of Equation 4 signifies the amount of information preserved after clustering while the second MI term restricts the compactness of the clusters. The Lagrangian parameter (β) is used to control the trade-off between these two terms.

To optimize the IB criterion in Equation 4, a greedy technique may be used. This translates to agglomerative clustering of the data. At each iteration, two clusters are merged such that the value of the objective function increases. This is given by:

$$\Delta\mathcal{F} = (p(C_i) + p(C_j))d_{i,j}, \tag{5}$$

where $d_{i,j}$ is the combination of Jensen-Shannon (JS) [13] divergence between two distributions:

$$d_{i,j} = JS(p(\mathbf{R}|C_i), p(\mathbf{R}|C_j)) - \frac{1}{\beta}JS((p(\mathbf{U}|C_i), p(\mathbf{U}|C_j)), \tag{6}$$

where the function JS is the Jensen-Shannon divergence measure between two probability distributions. The JS divergence between two probability distributions is the sum of the Kullback-Leibler divergences between the individual distributions and the average distribution [14].

As mentioned earlier, when applied to speaker diarization, the clusters are speaker segments (usually of length 2.5s). The relevance variables are posterior vectors obtained for each speech frame from a GMM estimated on the audio recording being diarized. Each element in the posterior vector is the posterior probability of a GMM component.

4.1. IB clustering algorithm

The IB based clustering algorithm, as used in speaker diarization task, is summarized below:

- **Input**: (a) data, (b) posterior probability of the relevance variable with respect to the input variable $\mathbf{U}$, (c) trade-off parameter (β).

- **Output**: Clusters $\mathbf{C} = \{C_1, C_2, \cdots, C_K\}$, where K is the number of desired clusters.
- Initialize the algorithm with each data point as its own cluster.
- Compute $\Delta\mathcal{F}(C_i, C_j)$ for all possible combinations of the clusters C_i and C_j.
- For l=1, $\cdots$, L (L is maximum number of clusters):
  - Merge the two closest clusters with the maximum $\Delta\mathcal{F}$.
  - Recompute the $\Delta\mathcal{F}$ between the new cluster and all other clusters.

The maximum number of clusters can be fixed in multiple ways. For instance, in speaker diarization a normalized MI based criterion is used. This is convenient as the MI terms are already computed.

4.2. IB for speaker clustering

In this section we adapt the IB clustering approach to obtain speaker labels on target domain data in the i-vector PLDA framework. The posteriors of relevance variables, input to the IB clustering algorithm as described above, need to be defined. A natural extension of the speaker diarization system would be to use posteriors of Gaussian components from the UBM. However, discarding completely the feature vectors of a recording and using only the posteriors of Gaussian components (from UBM) for speaker clustering task can be suboptimal. A better way to compute the posterior of the relevance variables is to use i-vector representation accounting for both the posteriors of Gaussian components (from UBM) and features vectors of an utterance. We build on the success of clustering i-vectors in [4] by deriving relevance variables based on i-vector cosine distances. We present two approaches based on the discussion above:

1. **Average zeroth order statistics:**
   Let $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T\}$ be an utterance with T number of feature vectors. Let us suppose that an UBM is also trained on these feature vectors with q^{th} Gaussian component of UBM represented by B_q. Use these Gaussian components of the UBM (B_q) as the relevance variables and the posterior probability of the relevance variables is computed by:
   $$p(B_q|\mathbf{O}) = \frac{1}{T}\sum_{t=1}^{t=T} p(B_q|\mathbf{o}_t). \quad (7)$$
   The probability $p(B_q|\mathbf{o}_t)$ can be computed from the parameters of the UBM. The quantity $p(B_q|\mathbf{O})$ is referred to as the average zeroth order statistics. The posterior probabilities of the relevance variables are computed for each of the utterances of target domain data. The IB clustering as described in Section 4.1 is used to obtain speaker labels with (a) input data being the utterances of target domain data ($\mathbf{O}$), and (b) posterior probability of relevance variable as the average zeroth order statistics ($p(B_q|\mathbf{O})$).
2. **Dot product:**
   The dot-product between i-vectors can be converted to probability scale and is used directly as the posterior probability of the relevance variable as described below:
   - Let $\mathbf{A}$ be the cosine distance measure matrix between the i-vectors in the target domain, where $(m,n)^{th}$ element of matrix $\mathbf{A}$ is given by:
     $$A_{m,n} = 1 - \frac{\mathbf{x}_m^t \mathbf{x}_n}{||\mathbf{x}_m||||\mathbf{x}_n||},$$
     which is the cosine distance between the i-vectors $\mathbf{x}_m$ and $\mathbf{x}_n$.
   - Convert each of these entries of matrix $\mathbf{A}$ into posterior probability of the relevance variable ($p_{m,n}$) using softmax function as given by the following equation:
     $$p_{m,n} = \frac{\exp^{A_{m,n}}}{\sum_n \exp^{A_{m,n}}}. \quad (8)$$
     The quantity $p_{m,n}$, referred to as posterior probability of the relevance variables, is computed for all the i-vectors. The IB clustering as described in Section 4.1 is used to obtain speaker labels with, (a) input data being the i-vectors of target domain data ($\mathbf{x}$), and (b) the posterior probability of relevance variable as the quantity $p_{m,n}$.

In this paper, we hypothesize that the speaker vectors in the PLDA space are more discriminative and thus will result in better clustering than the original i-vector space. The i-vector [3] projected in PLDA space (PLDA-vector) is expressed by the following equation:

$$\hat{\mathbf{y}} = (\mathbf{E} + \mathbf{V}^t\mathbf{\Sigma}\mathbf{V})^{-1}\mathbf{V}^t\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \quad (9)$$

The projected vector in the PLDA space is obtained for each i-vector and we refer to them as PLDA-vector.

Table 1. Performance of the i-vector based PLDA system on NIST-SRE 2010 male evaluation set by estimating the parameters in different datasets.

UBM and T-matrix	PLDA	EER (%)	minDCF
SWB	SWB	3.8	0.356
SWB	SRE	2.1	0.193

5. EXPERIMENTS

As defined in the the domain-adaptation challenge protocol[2], the Switchboard (SWB) and NIST Speaker Recognition Evaluation (SRE) datasets are used for system development. The Switchboard dataset is the source domain data and the domain adaptation protocol dictates that labels of this data are known. The SRE dataset is referred to as the target domain dataset as it matches the evaluation condition and labels of this dataset are unknown. The Switchboard data consists of 33,039 utterances with 3,114 speakers (both male and female). The SRE dataset contains 36,470 utterances drawn from the speaker evaluations dataset (SRE04 to SRE 08) with 3,500 speakers (male and female combined), out of which 13,628 utterances belong to the male set. The evaluation set is drawn from SRE 2010 evaluation condition 5, which is telephone enrolment and telephone test; the evaluation set consists of 3,465 target and 175,873 non-target male trials. In this paper, we report our results on the male set in terms of Equal Error Rate (EER) and minimum Decision Cost Function (minDCF) [15].

From Table 1, we observe that training the PLDA model with only the source domain data (SWB) results in 3.8% EER, whereas training the PLDA model with a labelled SRE development data gives 2.1%. Ideally, we want to reach the performance close to 2.1% EER (and 0.193 minDCF) using the unlabelled target domain data.

[2]http://www.clsp.jhu.edu/workshops/archive/ws13-summer-workshop/groups/spk-13/

Table 2. Performance of the baseline systems (speaker clustering based on cosine distance and PLDA scores) on NIST-SRE 2010 male evaluation set after domain adaptation.

System	EER (%)	minDCF
1 – i-vector	3.0	0.329
2 – PLDA-vector	2.9	0.323
3 – PLDA scores	2.9	0.316

Table 3. Performance of the proposed system (speaker clustering based on IB algorithm) on NIST-SRE 2010 male evaluation set. All systems perform better than the baseline systems in Table 2.

Relevance Variable	EER (%)	minDCF
Average zeroth order statistics	2.7	0.242
i-vector dot product	2.9	0.313
PLDA-vector dot product	2.5	0.225

5.1. Baseline system (Speaker clustering based on cosine distance and PLDA scores)

The performance of various baseline systems, namely (i) i-vectors with cosine distance (System 1), (ii) PLDA-vector with cosine distance (System 2), and (iii) PLDA Scores (System 3) is shown in Table 2. For System 1, the unlabelled target domain i-vectors are used for agglomerative clustering with cosine distance as metric. For System 2, the PLDA-vectors are obtained using Equation 9 and these PLDA-vectors are used for agglomerative clustering with cosine distance as the distance metric. As shown in Table 2, the performance of System 2 is marginally better than the first system in terms of minDCF. The best performance is obtained with 1,000 clusters for both the systems, which is close to the actual number of speakers in the target domain dataset (1,115 speakers). The best performance is obtained with interpolation parameter $\alpha \in [0.2, 0.3]$ suggesting that the contribution of the source domain parameters is greater than target domain PLDA parameters. System 3 was developed as follows: the distance between i-vectors was computed as in Equation 2 by using the source domain PLDA model parameters. This distance metric (PLDA score) is used for agglomerative clustering and the clustered output is used to train target domain PLDA model. The best performance is obtained with a α value of 0.2 and the performance of the system is better than System 1 and 2 in terms of minDCF.

5.2. Speaker clustering based on IB algorithm

The input to IB clustering algorithm as explained in Section 4, the posterior probabilities of the relevance variable, is critical in obtaining speaker labels. We explore three choices of posterior probabilities for speaker clustering:

- **Average zeroth order statistics**: As hitherto explained, we use average zeroth order statistics of an utterance as posterior probability of the relevance variable for clustering. From Table 3 (first row), we observe that it performs better than the baseline systems in terms of EER and minDCF. Thus the average zeroth order statistics of an utterance carries sufficient speaker discriminative information for IB clustering as evident from the results obtained.
- **i-vector dot product**: As described in Section 4, we convert the cosine distances between i-vectors to posterior probability, and IB clustering is used to obtain speaker labels. This system performs worse than the zeroth order statistics based system but still outperforms the baseline system in terms of minDCF. Although i-vectors are obtained from the average zeroth order statistics and feature vectors, the i-vector dot product system still performs worse than the average zeroth order statistics system on using IB method. Thus, the posterior probability of the relevance variable obtained by converting the dot product scores, as explained in Section 4, looses speaker discriminative information for clustering.
- **PLDA-vector dot product**: First we estimate the parameters of the source domain PLDA model. The i-vectors are projected in the source domain PLDA space to obtain PLDA-vectors as given by Equation 9. The posterior probability is computed as described in Section 4 by replacing the dot product between two i-vectors by dot product between two PLDA-vectors and IB clustering is used. It can be observed from Table 3 (third row) that this system performs the best in terms of EER and minDCF and is able to bridge the gap in performance up to 1.3% EER (from 3.8% to 2.5%). The PLDA-vectors are computed using a discriminative classifier (PLDA model) and hence IB method exploits the posterior probability of relevance variable using PLDA-vectors in an effective way.

6. CONCLUSIONS AND FUTURE WORK

We observed that training the PLDA model using labelled target domain data results in 45% reduction in EER compared to using only labelled source domain data (from absolute 3.8% to 2.1%). While assuming that the labels of target domain data is not known, we proposed to explore agglomerative clustering with different distance metrics to obtain speaker labels. The baseline system uses dot product distance metric for agglomerative clustering and it provides performance of 2.9% EER. Furthermore, we explored IB clustering technique (based on JS divergence metric) for obtaining speakers labels and found that it provides 14% relative improvement over the baseline system (from absolute 2.9% to 2.5%). We observed that the i-vector dot product system with IB clustering is the worst performing system which could be the effect of uncalibrated scores. Thus in future, we plan to perform calibration of scores obtained from dot product and subsequently derive the posterior probability of the relevance variables.

7. ACKNOWLEGEMENT

This work was supported by the project EU FP7 project EU Speaker Identification Integrated Project (SIIP).

8. REFERENCES

[1] Najim Dehak, Patrick J. Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[2] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, Mar 1992, vol. 1, pp. 517–520 vol.1.

[3] Daniel Garcia Romero and Alan McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 4047–4051.

[4] Daniel Garcia Romero, Alan McCree, Stephen Shum, Niko Brummer, and Carlos Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proceedings of Odyssey, Joensuu, Finland, 2014.*, 2014, pp. 260–264.

[5] Stephen H. Shum, Douglas A. Reynolds, Daniel Garcia-Romero, and Alan McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition," in *Proceedings of Odyssey, Joensuu, Finland, 2014.*, 2014, pp. 265–272.

[6] Elliot Singer, Douglas Reynolds, et al., "Domain mismatch compensation for speaker recognition using a library of whiteners," *Signal Processing Letters, IEEE*, vol. 22, no. 11, pp. 2000–2003, 2015.

[7] Hagai Aronowitz, "Inter dataset variability compensation for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 4002–4006.

[8] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, Nov 2003, pp. 411–416.

[9] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using hmm," in *IN PROCEEDINGS OF ICSLP-2002*, 2002, pp. 573–576.

[10] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard, "An information theoretic combination of MFCC and TDOA features for speaker diarization," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 2, pp. 431–438, 2011.

[11] Daniel Garcia Romero and Carol Y. Espy Wilson, "Analysis of ivector length normalization in speaker recognition systems," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27 to 31, 2011*, 2011, pp. 249–252.

[12] Niko Brümmer and Edward de Villiers, "The speaker partitioning problem," in *Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010*, 2010, p. 34.

[13] "Jensen-Shannon-Divergence," `https://en.wikipedia.org/wiki/Jensen-Shannon_divergence/`.

[14] Thomas M Cover and Joy A Thomas, *Elements of information theory*, John Wiley & Sons, 2012.

[15] "NIST Speaker Evaluation Recognition 2010," `http://www.nist.gov/itl/iad/mig/sre10.cfm`.