



Probabilistic Linear Discriminant Analysis for Robust Speaker Identification in Co-channel Speech

Navid Shokouhi, John H. L. Hansen*

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, TX 75080-3021, USA

{navid.shokouhi, john.hansen}@utdallas.edu

Abstract

Co-channel speech refers to a monophonic audio recording in which at least two speakers are present. Meeting and telephone conversations recorded on a single channel are examples of co-channel speech. In this study, we address the problem of speaker identification (SID) for trials that contain co-channel speech in the train and/or test sessions. The assumption here is that there is access to i-vectors for all the recordings and we would like to compensate for interfering speech without requiring any changes or enhancements on the audio. This is an attractive approach, since state-of-the-art SID systems are developed on i-vectors and thereby solutions that do not require alterations in the i-vector extraction stage are more convenient. We propose modifications to the standard PLDA formulation that enables extracting more accurate estimates of the eigenvoice matrix in the presence of interfering speech and consequently more accurate statistics for speaker dependent latent variables. The proposed *co-channel PLDA* formulation results in 30% relative drop in equal error rate when compared to the standard PLDA system for co-channel sessions with signal-to-interference ratios as low as 0dB.

Index Terms: speaker identification, co-channel speech, probabilistic linear discriminant analysis

1. Introduction

Co-channel speech refers to a monophonic audio recording in which at least two speakers are present. Mixing the channels from telephone conversations is an example of generating co-channel speech. Another instance would be meetings recorded with a single microphone. The presence of an interfering speaker dramatically drops the performance in most automatic speech applications and detecting/separating the undesired speech causes major difficulties. Unfortunately, interfering speech is fairly common in audio recordings. Those who followed Super Bowl XLIX may recall the post-game interview with Seattle Seahawks' head-coach, Peter "Pete" Carroll, where an audible interference from one of the broadcasting crew members made listening to Carroll's comments even more frustrating for Seahawks fans. That is a perfect example of a regular occurrence of co-channel speech, which we intend to account for in speaker identification (SID) in this study.

Speaker identification experiments can be highly influenced by the presence of secondary speakers, due to reduced reliability of the trained models. Although the target speaker is a com-

mon factor in all training samples for a given speaker model, the standard structure of SID systems has not been designed to average out interfering speech. Alternatively, automatically removing the interfering speakers from co-channel audio files is anything but practical for a large scale SID problem. Hence, a reasonable approach would be to focus on the model classification stage to remove the effect of co-channel speech. To the best of our knowledge, there are few studies that address SID in co-channel speech signals. In [1, 2], a description of the effects of artificially adding overlapped speech to train and test data in a Gaussian mixture model (GMM) based SID system is presented. There, the approach was to automatically detect and remove overlaps from co-channel speech [2]. Although many overlap detection algorithms have been investigated over the years [3, 4, 5, 6], none have considered solving the problem in the more general sense of co-channel interference. We differentiate co-channel speech from overlapped speech by considering the latter a special case of the former where both speakers are active at the same time. Co-channel speech refers to the broader case where the speakers are not necessarily overlapping (see Fig. 1). This definition disqualifies overlap detection solutions for the purposes of many large scale SID problems. It also gives rise to a more realistic problem, since only a small percentage of conversational speech may contain overlaps that can significantly drop SID performance [7, 5].

For the purposes of competitive SID performance, we focus our attention to i-vector systems that are considered state-of-the-art benchmarks for speaker verification. The introduction of i-vectors as low-dimensional, fixed-length feature vectors to represent the characteristics of audio sessions has vastly influenced speaker recognition [8]. In the standard i-vector speaker identification system, often a number of recordings are provided from which latent variables in the total variability space, aka i-vectors, are extracted from a high-dimensional model vector space. Using probabilistic linear discriminant analysis (PLDA) [9], these i-vectors can then be reduced to a secondary subspace for channel compensation [10, 11]. Recently, the NIST i-vector speaker recognition challenge introduced a task where participants were asked to perform speaker recognition on i-vectors provided by the organizers instead of audio recordings [12]. As a result of such popularities of i-vectors, it is therefore desirable to attempt to tackle co-channel interference in the i-vector level by devising a method for robustness against the presence of secondary speakers. One plausible scenario where it is likely to use this approach would be to consider SID tasks where the recordings provided for speaker verification are from co-channel conversations and we would like to verify the identity of target speakers without having to manually or automatically remove secondary speech from the

*This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

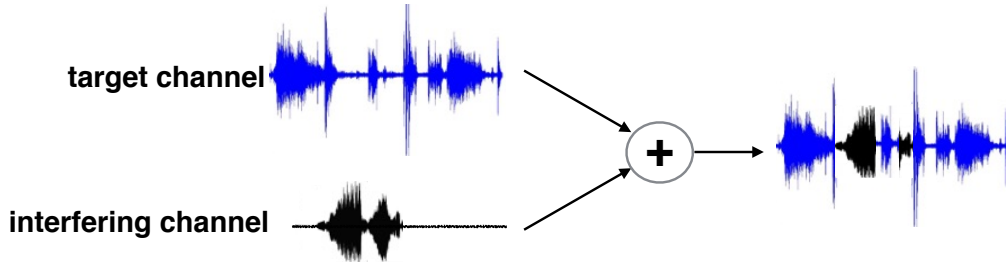


Figure 1: illustration of co-channel speech as defined in this study. In this context, co-channel and overlapped speech are not necessarily the same.

audio files before extracting i-vectors. In this case, the task of the SID system would be to also account for the fact that the i-vectors might be recorded from sessions that include undesired secondary speakers. There are many reasons why we would want to avoid enhancing/modifying the audio files to remove secondary speech; one being the fact that such algorithms also affect the target speaker and the background channel information [1, 13], while both speech and channel information should be kept unaltered for the purposes of SID.

The i-vector/PLDA solution is considered the pillar for recent SID systems [14, 15, 16, 17, 18]. PLDA uses inter-session and intra-session variabilities observed in several recordings in the development set corresponding to individual speakers to find a subspace in the i-vector space. The speaker-dependent latent variables in the resulting subspace have less channel dependency compared to the original i-vectors. The aim of this study is to find a modified version of the PLDA paradigm to make i-vectors collected from co-channel sessions usable for SID experiments and create overall robustness with respect to interfering speech. Here we investigate the possibility of performing an i-vector normalization strategy to tackle co-channel interference. An important aspect of our proposed method would be to find the least amount of information and data required to improve co-channel SID. In other words, it is important to us that the experiments be easily replicable in large-scale SID evaluations. This is partly accomplished by mixing the channels from telephone conversations in Switchboard and NIST SRE data. This allows full control over the average signal-to-interference ratio (SIR) of background development data, without requiring additional data specifically collected for the purposes of this study.

We start by briefly describing the standard PLDA approach for channel compensation and the modifications that have led to the most common state-of-the-art system [9, 10]. This is followed by our description of the co-channel speech condition for SID and our proposed *co-channel PLDA* formulation through which we intend to remove speaker interference. Section 3 presents an illustration of the system setup and an evaluation of the co-channel PLDA method. Our current work is concluded in the final section and given alongside an outline of our future plans to further improve this study.

2. Probabilistic Linear Discriminant Analysis

The PLDA algorithm referred to in this study was introduced in [9] for face recognition. It was later applied to speaker recognition as a channel compensation method in [10] with some modifications to better suit the problem. Nowadays, PLDA is considered one of the standard supervised approaches in dealing with channel mismatch. The work in this study excludes

descriptions of the i-vector extraction and focuses on the post processing of i-vectors that takes place in PLDA.

2.1. Standard PLDA:

Given n_i observation i-vectors for speaker i from a set of development speakers, the PLDA model assumes the following linear factorization for each i-vector m_{ij} :

$$m_{ij} = \mathbf{V}y_i + \mathbf{U}x_{ij} + z_{ij}, \quad j = 1, \dots, n_i \quad (1)$$

where the speaker and session dependent latent variables, y_i and x_{ij} , take a standard normal distribution, $\mathcal{N}(0, \mathbf{I})$. \mathbf{V} and \mathbf{U} are typically tall matrices that represent the eigenvoice and eigenchannel subspaces, respectively. The session-dependent slack variable, z_{ij} is normally distributed with a diagonal covariance matrix, $\mathcal{N}(0, \mathbf{\Sigma})$, [10, 9]. PLDA predicts the model parameters, $(\mathbf{V}, \mathbf{U}, \mathbf{\Sigma})$, using the expectation-maximization (EM) algorithm [9]. After estimating the subspace characteristics using background development data, trial components are then reduced to the same speaker-dependent subspace (indirectly) using the PLDA model parameters and scored through a hypothesis testing recognition procedure (see [9] for details).

2.2. Simplified PLDA:

The simplified PLDA framework for SID omits the eigenchannel term by including all non-speaker dependencies into the slack variable, z_{ij} . This is done by assuming a full covariance matrix for z_{ij} . This method is more desirable partly due to the fewer degrees of freedom in estimating the parameters [19]. Equation (2) shows the simplified PLDA formulation.

$$m_{ij} = \mathbf{V}_{ij}y_i + z_{ij}, \quad (2)$$

where $z_{ij} \sim \mathcal{N}(0, \mathbf{\Sigma}_f)$ and $\mathbf{\Sigma}_f$ is a full covariance matrix. The variable z_{ij} is sufficient to characterise channel dependencies. Using a full covariance matrix brings redundancy to the use of an eigenchannel matrix.

2.3. Co-channel PLDA (proposed method):

When considering i-vectors extracted from co-channel recordings, interfering speech adds an additional kind of variability to the i-vector space. We address this problem by considering two tactical approaches:

1. Constructing co-channel background data to train the PLDA models.
2. Modifying the PLDA formulation to fit the co-channel speech paradigm.

The first approach relies on the same ability of the PLDA framework that compensates channel mismatch. PLDA recognizes the variabilities observed across different recordings for

a given speaker and removes them from the speaker subspace (aka the eigenvoice factors in \mathbf{V}). A natural approach for co-channel speech in this case would be to treat interfering speech as channel mismatch. This can be accomplished by adding co-channel i-vectors to the PLDA background data and leaving it to the PLDA model estimation process to recognize interfering speech and remove its effects in the speaker subspace. However, in this scenario interfering speech is not effectively picked up by neither the eigenchannel matrix (in case of the standard PLDA in (1)) nor the slack variable (in case of (2)). We address this by adding our second approach which is to add a speaker dependent term intended to model interfering speech. Since this term also corresponds to speaker identities, it should as well be represented by the speaker subspace. The resulting factorization is:

$$m_{ij} = \mathbf{V}y_i + \mathbf{V}'w_{ij} + z_{ij}. \quad (3)$$

Since the model still needs to represent channel variabilities, we keep z_{ij} as a full covariance normal distributed vector. \mathbf{V}' represents the speaker dependent subspace, as does \mathbf{V} , with the difference that one is a rotation of the other with an unknown rotation factor. w_{ij} is the latent variable which corresponds to the interfering speaker. There are a few reasons that prevent us from directly using \mathbf{V} as factor loadings for the interfering components in (3), one being that as part of the degrees of freedom of the PLDA solution, the eigenvoice matrix can only be determined up to an unknown rotation matrix [19] (similarly for \mathbf{V}'). Using different notation reminds us of this limitation. This prevents us from directly using \mathbf{V} to represent the interfering speaker term. However, since we know that \mathbf{V} and \mathbf{V}' must be related by a rotation matrix, we can use this knowledge to simultaneously update \mathbf{V} and \mathbf{V}' in the EM iterations.

As discussed, the relation between \mathbf{V} and \mathbf{V}' is characterized by an unknown rotation matrix, \mathbf{R} :

$$\mathbf{V} = \mathbf{R}\mathbf{V}'. \quad (4)$$

A reasonable \mathbf{R} can be estimated via singular value decomposition (SVD) by considering the columns of \mathbf{V} and \mathbf{V}' as data points in the speaker dependent subspace. The rotation matrix is derived from the cross-variance between \mathbf{V} and \mathbf{V}' basis functions, \mathbf{S} :

$$\mathbf{S} = \sum_{i=1}^{N_V} \tilde{v}_i \tilde{v}_i^T, \quad (5)$$

where \tilde{v}_i is v_i are the eigenvoice basis vectors centered at the origin and N_V is the number of columns in \mathbf{V} (and/or \mathbf{V}' , since both have the same dimensions). The rotation matrix is defined as below:

$$\mathbf{R} = \mathbf{S}_{row} \mathbf{S}_{col}^T, \quad (6)$$

where \mathbf{S}_{col} and \mathbf{S}_{row} are the column and row spaces of \mathbf{S} obtained from SVD:

$$\mathbf{S} = \mathbf{S}_{col} \mathbf{\Lambda} \mathbf{S}_{row}^T. \quad (7)$$

The rotation matrix is used in each iteration of the EM algorithm to update the matrix \mathbf{V}' and align it with \mathbf{V} ,

$$\mathbf{V}' \leftarrow \mathbf{R}\mathbf{V}'. \quad (8)$$

Updating \mathbf{V}' before estimating the statistical statistics of latent variables, y_i and w_{ij} , removes the redundancy in the second term of equation (3) by replacing the eigenvoice matrix \mathbf{V}'

with information obtained from the basis vectors in \mathbf{V} . Since both factors in (3) are guided towards representing the speaker space, the overall system achieves a better estimate of \mathbf{V} and consequently more accurate estimates for the statistics of the hidden variables.

3. Experiments and results

This section presents the experimental setup through which the proposed *co-channel PLDA* method is evaluated. First, we describe system specifications and the method of generating co-channel sessions. Next, the results obtained by applying the PLDA alternatives (from Sect. 2) are presented.

3.1. Co-channel SID in Switchboard:

Our speaker identification experiments are performed on over 300 hours of Switchboard II Phase 2 recordings. The universal background model (UBM) and total variability (TV) matrix are estimated from 450 hours of single channel data picked from NIST SRE 04,05, and 06 [20, 21, 22]. PLDA background data is generated from over 280 hours consisting of approximately 220 target speakers from the NIST SRE 10 [23] telephone channels and phone-call recording style sessions (tel-phn). Using telephone recordings guarantees two single-speaker channels per session. This gives us control over the average SIR values of the co-channel files which are generated by mixing the two channels provided for each 5 minute session. In the PLDA background, when required, the target and interfering speakers' speech are mixed according to a desired SIR value for each session. For example, in cases where we use clean speech (w/o co-channel interference) to train the PLDA model parameters, the SIR is fixed at 100dB by setting the gain for the secondary speaker to a low enough value, keeping the secondary speaker's presence almost completely unnoticeable. In cases where we introduce co-channel interference, prior to mixing the channels, the overall energy of the interfering channel is reduced by a gain factor to achieve the desired SIR. The same procedure is used to generate co-channel speaker models for trials using the two channels available for each Switchboard session. The "mixing" of signals is the act of summing the two channels, as illustrated in Fig. 1. The target speaker was the common factor between all co-channel files designated for each speaker model in both the model training set and the PLDA background data.

Features used in our experiments are 39 dimensional; consisting of 13 dimensional MFCC vectors and their corresponding first and second order differences (Δ and $\Delta\Delta$). The i-vector/PLDA system configuration for baseline experiments uses 1024 mixtures for the UBM, 400 dimensional i-vectors, and the PLDA assumes a 200 dimensional subspace for the eigenvoice latent components. Prior to applying PLDA, i-vectors are reduced to 220 dimensional vectors using linear discriminant analysis (LDA).

We construct 4 different trial sets, each containing co-channel files for training speaker models from one of the SIR levels: 100dB, 10dB, 5dB, 0dB. All results presented here are from SID experiments on male speakers. The trial sets are used to compare the following three systems:

- *clean PLDA*: The simplified PLDA model described in (2) trained on a clean development set (without co-channel interference).
- *mixed PLDA*: Uses the same model as in *clean PLDA*, with the difference that it is trained on a development

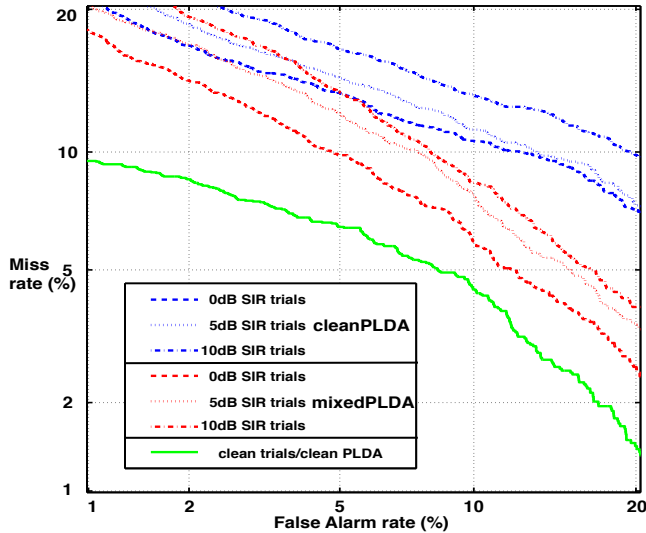


Figure 2: Comparing DET curves before and after including co-channel speech in the PLDA background data (aka clean (blue) vs. mixed (red) PLDA) across different SIR values in the trials. The performance of under clean conditions (no co-channel in the trials) is provided in green as a point of reference.

set that contains both clean sessions and 0dB SIR co-channel sessions.

- *cch PLDA*: Short for *co-channel PLDA*, this system uses the development set of *mixed PLDA* and is trained using the proposed model defined in (3).

3.2. clean vs. mixed PLDA:

Results show that SID error rates in co-channel trials reduce when co-channel sessions are introduced to the PLDA background data, as opposed to the regular PLDA setup where all files are free of co-channel interference, which we call “clean” PLDA. The approach in *mixed PLDA* is to replace a subset of each speaker’s recordings with co-channel sessions (as described in Sect. 3.1). In our experiments, we replace half of the clean PLDA training data with co-channel sessions. Figure 2 compares the detection error trade off (DET) curves for *clean* and *mixed PLDA*. The performance is shown for three different SIR values.

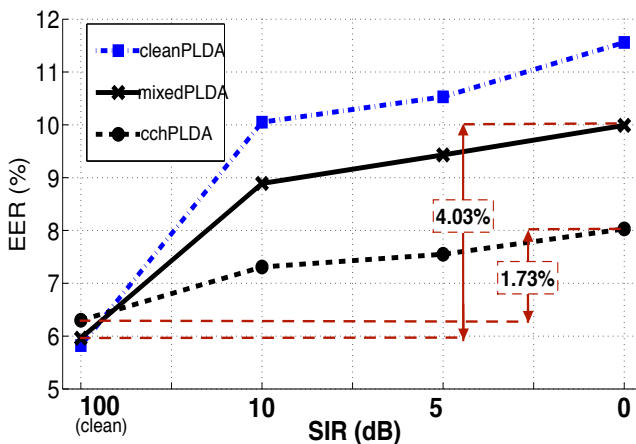


Figure 3: EER comparison for 3 systems: *clean PLDA*, *mixed PLDA*, and *co-channel (cch) PLDA* (proposed method). Across four SIR values: 100dB(clean), 10dB, 5dB, 0dB.

This approach takes advantage of the channel compensation concept designed in the PLDA formulation. Although effective to some extent, it neglects the fact that in co-channel speech the interference and the target signal are of the same kind and should be defined in the same subspace. This is addressed in the next section, where we present the results for *co-channel PLDA*.

All DET curves assume equal coefficients for the miss rate and false-alarm rate ($C_{miss} = C_{fa} = 1$) and a target to impostor ratio of 0.001.

3.3. co-channel PLDA:

As described in Sect. 2.3, *co-channel PLDA* attempts to model a linear factorization of the i-vectors into a target speaker and an interfering speaker component (see Eq. (3)). PLDA is able to distinguish the target speaker using the several recordings available for each speaker. The key difference between this method and *mixed PLDA* is that the system is forced to use a similar subspace to model the interfering speaker. Figure 3 shows the equal error rate (EER) for different amounts of co-channel interference introduced to the trials.

Considering that *clean PLDA* does not claim robustness towards co-channel interference, it shows little resistance as trial SIR values increase. Since *mixed PLDA* has some observation of the co-channel condition, it reduces the EER for at least 1% across all SIR conditions. *Co-channel PLDA*, further improves the performance and obtains 2.5-3.5% drop in EER in all conditions. EER variations are also significantly different for the *co-channel PLDA* model compared to the other two systems. Figure 3 shows that the clean-to-0dB EER range for *mixed PLDA* is more than twice as much as *cch PLDA*.

4. Conclusions

A modified formulation of the PLDA framework for speaker identification was proposed to address co-channel speech interference for speaker verification tasks. The proposed method (co-channel PLDA), is designed to find a linear factorization of the i-vectors into a target and an interfering speaker component by recognizing that the two components should belong to the same subspace. This is done by adjusting the the interference factor loadings to align eigenvoice subspace. Results show that the proposed method relatively reduces the EER by as much as 30% for 0dB co-channel interference in the trial set, when compared to the standard PLDA formulation. The absolute dynamic range of the EER across clean-to-0dB SIR trials is 1.73% for co-channel PLDA and 4.03% for the standard PLDA. For our future work, we are interested in investigating the effects of male vs. female co-channel interference. It would also be useful to consider the same approach on meeting style co-channel interference, which will require reliable SIR estimations.

5. References

- [1] R. E. Yantorno, “Cochannel speech study,” Electrical and Computer Engineering Department Temple University, Tech. Rep., September 1999.
- [2] R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, “Effects of co-channel speech on speaker identification,” in *SPIE Intl. Symp. on Tech. for Law Enforcement*, November 2000.
- [3] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, “Overlapped speech detection for improved diarization in multiparty meetings,” in *Proc. ICASSP*, Las Vegas, Nevada, 2008, pp. 4353–4356.
- [4] N. Shokouhi, A. Sathyanarayana, S. Sadjadi, and J. H. L. Hansen, “Overlapped-speech detection with applications to driver assess-

- ment for in-vehicle active safety systems,” in *Proc. IEEE ICASSP*, Vancouver, BC, May 2013.
- [5] B. Smolenski and R. Ramachandran, “Usable speech processing: A filterless approach in the presence of interference,” *Circuits and Systems Magazine, IEEE*, vol. 11, no. 2, pp. 8–22, 2011.
- [6] K. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, “Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions,” in *IEEE Intl. Symp. on Intelligent Signal Processing and Communication Systems, ISPACS*, November 2000, pp. 710–713.
- [7] O. Cetin and E. Striberg, “Speaker overlaps and asr errors in meetings: Effects before, during, and after the overlap,” in *Proc. IEEE ICASSP-2006: Int. Conf. Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, pp. 357–360.
- [8] 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.
- [9] S. Prince and J. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, Oct 2007, pp. 1–8.
- [10] P. Kenny, “Bayesian speaker verification with heavy-tailed priors,” in *Proc. ISCA Odyssey - The Speaker and Language Recognition Workshop*, 2010.
- [11] D. Garcia-Romero and C. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” in *Proc. INTERSPEECH*, Florence, Italy, Sept. 2011, pp. 249–252.
- [12] C. S. Greenberg, D. Bans, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, “The nist 2014 speaker recognition i-vector machine learning,” in *Proc. ISCA Odyssey*, Singapore, Singapore, Jun. 2012.
- [13] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, “Co-channel speaker separation by harmonic enhancement and suppression,” *IEEE Trans. on Speech and Audio Process.*, vol. 5, no. 5, pp. 407–424, September 1997.
- [14] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, “Full-covariance ubm and heavy-tailed plda in i-vector speaker verification,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 4828–4831.
- [15] S. Cumani, O. Plchot, and P. Laface, “Probabilistic linear discriminant analysis of i-vector posterior distributions,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 7644–7648.
- [16] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brummer, “Discriminatively trained probabilistic linear discriminant analysis for speaker verification,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 4832–4835.
- [17] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, “Towards noise-robust speaker recognition using probabilistic linear discriminant analysis,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 4253–4256.
- [18] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, “Full-covariance ubm and heavy-tailed plda in i-vector speaker verification,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 4828–4831.
- [19] A. Sizov, K. Lee, and T. Kinnunen, “Unifying probabilistic linear discriminant analysis variants in biometric authentication,” in *Proc. S+SSPR*, 2014.
- [20] NIST, “The NIST year 2004 speaker recognition evaluation plan,” 2008. [Online]. Available: <http://www.nist.gov>
- [21] —, “The NIST year 2005 speaker recognition evaluation plan,” 2008. [Online]. Available: <http://www.nist.gov>
- [22] —, “The NIST year 2006 speaker recognition evaluation plan,” 2008. [Online]. Available: <http://www.nist.gov>
- [23] —, “The NIST year 2010 speaker recognition evaluation plan,” 2010. [Online]. Available: <http://www.nist.gov>