Robust Speaker Recognition in Co-channel Speech Using Probabilistic Linear Discriminant Analysis

Navid Shokouhi, *Student Member, IEEE,* John H. L. Hansen, *Fellow, IEEE*

***Abstract*—The abstract goes here.**

***Index Terms*—Co-channel speech, speaker recognition, probabilistic linear discriminant analysis**

I. Introduction

Co-channel speech refers to single-channel audio signals that contain more than one speaker. In this study, we address the problem of speaker recognition for co-channel recordings. Co-channel speech refers to single-channel audio signals that contain more than one speaker. The presence of speech interference is an important artifact for all automatic speech processing systems. As speech technology continues to advance, the need to address multi-speaker interference increases.

The difference between what we present here and every other study in this area is that we would like to bypass solutions that require removing interfering speech from the original signal, which are primarily known as speaker diarization. Speaker diarization is defined as the task of determining "who spoke when?" within an audio recording. Speaker recognition in co-channel speech using diarization as a preprocessing step involves recognizing each speaker within the recording, which is itself a speaker recognition problem. One can easily see the inherent logical issue in this approach; where in order to achieve $\mathbf{A}$, $\mathbf{B}$ is required. And performing $\mathbf{B}$ requires some variation of $\mathbf{A}$. Alternatively, we are interested in modifying model parameters extracted from co-channel data in a way that would only represent the primary speaker. Currently, the most common parameterization of speaker dependent models is in the form of i-vectors [**?**]. I-vectors are latent parameters that model the covariance of speaker/session-dependent Gaussian mixture models (GMM) with respect to a generic GMM (aka Universal background model). The UBM is ideally both session- and speaker-independent. The use of i-vectors, has become a standard way of modeling speaker specific traits for speaker recognition. In a way that in many cases i-vector extraction is considered a preprocessing step in performing speaker recognition on a data-set. Therefore, it is both reasonable and desirable to concentrate on post i-vector analysis to deal with co-channel speech interference [1]. The goal of this study is to build upon the latent variable perspective, popularized by i-vectors [**?**] and its predecessors [**?**], to improve speaker recognition in co-channel signals. This provides the luxury of short-circuiting speaker diarization, which in addition to what we described above is also a computationally intensive solution.

Speaker recognition experiments can be highly influenced

Mail All Correspondence To:

Prof. John H.L. Hansen
Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Dept. of Electrical Engineering, University of Texas at Dallas
2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, U.S.A

*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.

by the presence of secondary speakers, due to reduced reliability of the trained models. Although the target speaker is a common factor in all training samples for a given speaker model, the standard structure of speaker recognition systems has not been designed to average out interfering speech. To the best of our knowledge, there are few studies that address speaker recognition in co-channel speech signals. In [2], [3], a description of the effects of artificially adding overlapped speech to train and test data in a Gaussian mixture model (GMM) based system is presented. There, the approach was to automatically detect and remove overlaps from co-channel speech [3]. Although many overlap detection algorithms have been investigated over the years [4], [5], [6], [7], none have considered solving the problem in the more general case of co-channel interference. This study differentiates 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. **??**). This study focuses on speaker recognition in co-channel speech interference, in which overlaps may occur. This definition disqualifies overlap detection solutions for the purposes of many large scale speaker recognition problems. It also sheds light on a more realistic problem, since only a small percentage of conversational speech contains amounts of overlap that are large enough to significantly drop speaker recognition performance [8], [6].

To further illustrate our problem statement, we set the following ground rules. We assume that:

- There is sufficient data from multiple recording sessions for each speaker.
- Recordings are co-channel and contain data from speakers other than the person of interest.

In the standard i-vector speaker recognition framework, often a number of recordings are provided for each speaker. Using probabilistic linear discriminant analysis (PLDA) [9], these i-vectors can then be reduced to a secondary subspace to compensate channel variations[1] across different recordings [10], [11]. Therefor, the latent variables in the PLDA subspace are calculated in a way to only represent speaker-dependent information [12], [13], [14], [15], [16]. Now if i-vectors are to be extracted from co-channel signals, the speaker-dependent latent variables from PLDA represent a combination of all speakers in the original audio file. In the case of speaker recognition in co-channel speech, the task of our proposed system would be to also account for the fact that i-vectors might have been extracted from co-channel sessions.

The aim of this study is to develop a modified version of the PLDA paradigm to make i-vectors collected from co-channel sessions usable for speaker recognition experiments and create overall robustness with respect to interfering speech. PLDA uses inter-session and intra-session variabilities from a development set to find a subspace in the i-vector space that best represents speaker dependencies. Here we investigate the possibility of performing an i-vector normalization strategy to by considering co-channel interference a form of inter-session variability. It is important to us that our experiments be easy to replicate and require minimal meta-data (labels, speaker and channel information, etc.).

An investigative approach to the effects of co-channel speech in speaker recognition is presented in the next section, Sect. II. We lay a groundwork by showing performance drop caused by adding co-channel with different signal-to-interference ratio (SIR) to enrollment and test data. In Sect. III, standard PLDA and its preceding modified version, simplified PLDA, are described. We will state how channel compensation is performed through these methods [9], [10]. Section III also investigates treating co-channel interference in a manner

[1]Channel mismatch refers to differences in recording conditions and devices. The authors feel obligated to remind readers not to confuse channel information with co-channel speech.

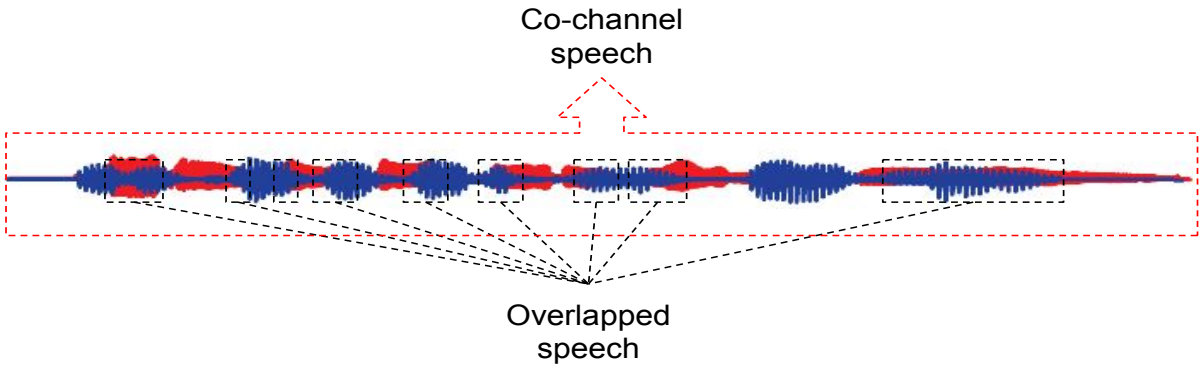

Fig. 1: *Difference between co-channel and overlapped speech. Overlap refers to instances where more than one speaker is active. Co-channel is defined as an entire stream that contains multiple speakers. All co-channel files do not necessarily contain overlap.*

similar to how PLDA addresses channel mismatch, using a background data preparation scheme we call *mixed PLDA*. This will be followed by our proposed *co-channel PLDA* formulation, which is used to remove speaker interference (Sect. **??**) from PLDA's speaker-dependent latent variable subspace. Section **??** illustrates practical simplifications to our proposed model and analyzes its convergence behavior. Section **??** describes our experimental framework and presents results co-channel PLDA results.

II. Effect of Co-channel in Speaker Verification

The first step to addressing the problem of co-channel speech compensation in speaker verification is to establish a quantative perspective as to how much performance degradation one should expect. There have been a number of studies that have investigated "co-channel speech" and its effect on speaker verification. However, each provides different insight due to the somewhat nuanced definition of co-channel, as explained in the introduction. Many consider overlap synonymous to co-channel, which we strongly argue against in this study. In [**?**], co-channel speech in the form of overlaps significantly increases equal error rates for speaker verification systems based on Gaussian mixture models (GMM). An interesting result presented in [**?**] is that a strict removal of all overlaps does not yield optimal performance, but rather keeping all "usable speech" leads to best performance under co-channel. Therefore, usable speech detection was proposed instead overlap detection to improve speaker verification performance, which has become a standard approach to addressing overlapped speech in speaker verification [**?**], [**?**]. In a sense, usable speech refers to speech from the foreground speaker (speaker of interest) with high signal-to-interference ratio and/or all voiced segments of the foreground speaker in which spectral harmonic patterns have not been severely disrupted [**?**]. Another analytic study on overlap in speaker verification was presented in [**?**], where authors show that adding overlap to test data results in more performance degradation compared to train data. The authors argue that an averaging effect occurs when multiple instances of overlapped training data is provided in enrollment sessions, while test data usually has a more direct role in deriving likelihood ratios for each trial. An alternative to removing overlapped segments for speaker recognition has been to perform speaker separation [**?**], [**?**], or in some cases simultaneous identification of both speakers in an overlapped stream [**?**], [**?**]. We see the main theme in many of these studies, which has been to focus on overlapped speech, rather than the more general case of co-channel speech. Although overlap presents an undeniably difficult challenge in speaker verification, one can argue that the amount of overlap in conversational co-channel speech is far too small to pose a significant threat to speaker verification performance in large scale verification problems such as NIST

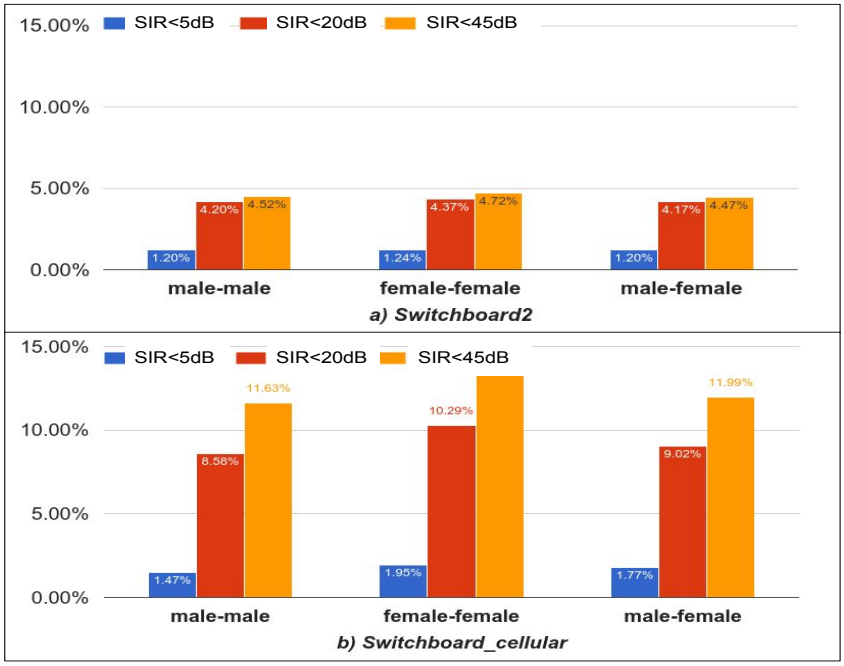

Fig. 2: *Percentage of overlaps to total speech in Switchboard2 and Switchboard cellular telephone conversations. Three SIR upper bounds are selected to label overlaps; 5dB, 20dB, 45dB. The higher the SIR upper bound, the stricter the overlap labels. Separate results are shown for male-male, male-female, and female-female conversations.*

speaker recognition evaluations, where in many cases sufficient data is provided. Later in this section, we will separately evaluate system performance under overlap-only conditions. But first, a useful analysis would be to see exactly what percentage of everyday conversational data contains overlaps. Readers are encouraged to visit [**?**] for a detailed analysis of overlaps in conversational speech corpora.

Three popular corpora are investigated here:

- Switchboard2: a large collection of five minute telephone conversations involving several hundred speaker from across the United States.
- Switchboard Cellular: five-six minute telephone conversations on cellular phones.
- AMI meeting corpus: A dataset consisting of 100 hours of meeting recordings from several locations across Europe.

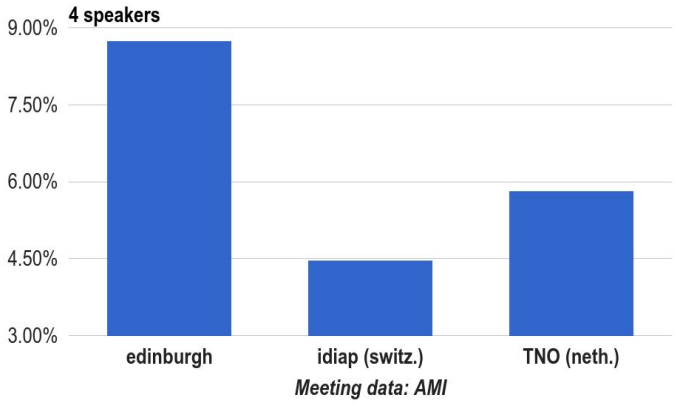

Fig. 3: *Percentage of overlaps to total speech in the AMI meeting corpus. All meetings used here have exactly 4 speakers. The percentage of overlap is significantly higher here compared to Switchboard (compare with blue bars in Fig. 2).*

To estimate overlap in Switchboard2 and Switchboard Cellular, the separate (almost interference-free) channels provided for each speaker are first segmented into speech and silence using an energy based voice activity detection. Signal energies are required at each time-sample, since signal-to-interference ratio (SIR) is used to define overlap. Note that we use absolute SIR, since none of the speakers are favored in this scenario.

$$SIR(n) = |10log_{10}(\frac{P_1(n)}{P_2(n)})| \quad (1)$$

where $P_1(n)$ is the per-sample energy of channel 1 and $P_2(n)$ corresponds to channel 2.

The channels are mixed (per sample addition of the two signals) to create co-channel data. Instances at which both

speakers are active are considered overlapped. Since the SIR value varies for different segments, overlaps are thresholded. Segments with SIR lower than the threshold are considered overlapped. The amount of overlap varies with the maximum allowable SIR set by the evaluator. For instance, one might consider the mere presence of two speakers at the same time sufficient to label a segment as overlap, which is an indication of high SIR thresholds ($45dB$ in Fig. 2). A more pragmatic view, however, is to choose an SIR small enough to preserve as much data possible for the sake of not loosing speaker recognition performance as a result of less data. This pragmatic view is shared in many studies under the definition of usable speech [**?**]. Of course lower SIR values, $(0-5)dB$, have a more significant impact on speaker verification. But to provide more insight to the reader, three SIR upper bounds have been used in Fig. 2; 5, 20, $45dB$. We argue that any overlap up to $5dB$ should have noticeable impact on speaker verification. An upper bound of $45dB$ is also chosen, since the percentage of overlap converges beyond $45dB$. We see that with a $5dB$ threshold, the percentage of overlap to total speech is below 2%. For the curios reader, we have provided overlap percentages for three groups of conversations: male-male, male-female, and female-female pairs. It is clear that gender does not play a role in dictating overlap percentage, at least in Switchboard.

For a broader perspective, we also analyze the AMI meeting corpus. The difference between meetings and phone conversations is in the number of speakers and face-to-face interaction. we speculate that number of speakers increases overlap while on the other hand the fact that all speakers are present in the same room (i.e., face-to-face interaction) limits the amount of overlap. Another difference is that SIR is not as well defined for meetings as it is for two-party phone-calls, since multiple parties may be active at the same time. In Fig. 3, we assume at least two speakers should have a relative SIR of up to $5dB$. Any additional speaker is evaluated with respect to the primary speaker, but with a $20dB$ threshold. As Fig. 3 suggests, location also plays a significant role in overlap percentage. The authors refrain from speculating the impact of location, since it exceed the scope of this study.

In text-independent speaker recognition, where we are interested in long-term acoustic characteristics, 4-5$\%$ of overlapped speech in our data has little effect on speaker recognition accuracy. The point here is not to say that overlapped speech can be neglected in speaker recognition, but to clarify that speaker recognition in its most common form is more concerned with co-channel speech in general rather than *overlap* as it appears in everyday English conversations. In the more general case of co-channel speech interference, the presence of secondary speakers has a significant impact on speaker verification. We roughly estimate that in a two-party phone conversation, approximately 50% of the data contains the unwanted secondary speaker (compare this with 2% in Fig. 2).

A. Co-channel Interference in Trials

In this segment, a series of speaker verification experiments are conducted on Switchboard2 to examine the effect of introducing co-channel speech in enrollment and test data. Single-speaker data is used to estimate PLDA parameters in this section (rather than co-channel data), since our intention is to show performance drop due to co-channel speech in trials in a typical i-vector/PLDA system. Trials are evaluated at different levels of co-channel interference (i.e., SIR level). In each scenario, trial recordings are summed with their counterpart channel from the phone conversation to create co-channel data, as if speakers are speaking on a single channel (shown in Fig. 4). Speaker labels for trial recordings are generated based on the foreground speaker. In this context, foreground speaker refers to the speaker of interest. For example, in a $5dB$ co-channel session generated from Switchboard2 containing

speakers **X** and **Y**, if **X** were the foreground speaker, the average energy of **X** would be $5dB$ higher than the average energy of **Y**. Therefore, SIR_{trial} is defined slightly differently from here on after, compared to the per-sample absolute $SIR(n)$ in (1).

$$SIR_{trial} = 10log_{10}\Big(\frac{\frac{1}{T}\sum_t E_X(t)}{\frac{1}{T}\sum_t E_Y(t)}\Big), \qquad t = 1, ..., T \quad (2)$$

where $E_X(t)$ and $E_Y(t)$ are signal energies at frame t. Five SIR levels are chosen throughout experiments; $100dB$ (i.e., clean sessions), $20dB$, $10dB$, $5dB$, and $0dB$. In $0dB$ the average energy of the foreground and background data is equal. To avoid mismatch, the clean condition is also generated through the same procedure with an SIR of $100dB$ favoring the foreground speaker.

A gender-independent universal background model (UBM) is created using 8kHz single-speaker NIST SRE data from 2004, 2005, and 2006 challenges [17], [18], [19]. The UBM consists of 2048 Gaussian mixtures representing a 39 dimensional feature space (13 dimensional MFCC plus Δ and $\Delta\Delta$). The same data from SRE 2004-6 is used to estimate a total variability (TV) matrix, which extracts 400 dimensional i-vectors [20]. The data used here to estimate PLDA parameters are single-speaker recordings from NIST SRE 2008 [21]. PLDA training data consists of approximately $11k$ single-speaker utterances from over 1300 speakers. Trial data is developed from 2500 Switchboard2 recording sessions containing approximately 800 speakers. Prior to feature extraction, trials are processed using comboSAD, an unsupervised speech activity detection [22], which has previously shown to provide stable performance improvement in such speaker recognition tasks [23].

Figure 5 shows speaker verification performance for the five SIR cases. As shown, EER for the clean condition is significantly lower than all the other SIR levels, even $20dB$. The sudden jump in EER shows the significance of co-channel interference.

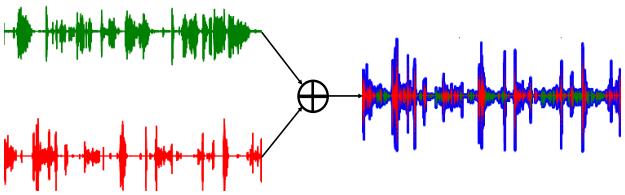

Fig. 4: *Mixing two channels of a Switchboard phonecall. The example here mixes the signals with 0dB SIR. Blue shows the resulting co-channel signal. Red and green each show one of the single-speaker signals.*

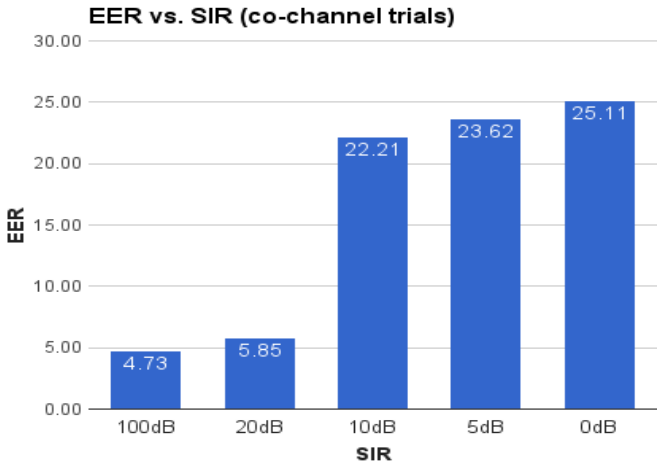

Fig. 5: *Speaker verification performance with co-channel speech in switchboard trials. The i-vector/PLDA system uses a typical system configuration and is fully trained on single-speaker data. The purpose of this chart is to show the rapid increase in equal error rate (EER) as co-channel data is added to the trials. 100dB SIR represents clean (single-speaker) trials.*

To see how much performance drop is due to *overlap*, another experiment is conducted, where all speech from the secondary speaker is dropped from the recordings, except for segments that overlap the foreground speaker. This is accomplished by using voice activity detection (VAD) labels from the 100dB trials, while using $0dB$ audio data for the trials. Figure 6, compares speaker verification under overlap with 0dB co-channel speech.

III. Motivation

This section provides a brief overview of the chronological introduction and development of probabilistic linear discriminant analysis (PLDA) in speaker recognition. PLDA was initially proposed for face recognition in [9]. It was later adopted as a channel compensation step for speaker recognition using

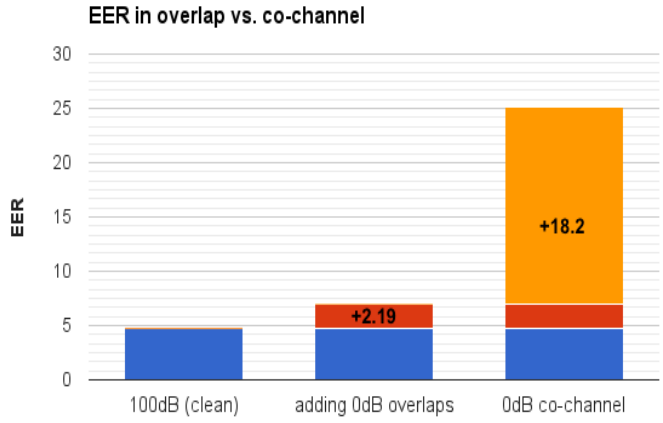

Fig. 6: *Comparing the effect of overlap in speaker verification with the more general case of co-channel. This study differentiates overlap from co-channel speech by considering overlaps to be segments during which both speakers are active. Co-channel refers to the more general case of two speakers in an audio stream, not necessarily overlapped (see Fig. 1). The chart shows that overlap plays a small part in the rise of EER compared to co-channel interference.*

i-vectors [**?**]. A number of studies since then have presented different formulations for the factor loading paradigm most commonly known as PLDA. In this study, we are specifically interested in three formulations: 1) standard PLDA [10], 2) simplified PLDA [**?**], [**?**], and 3) the two-covariance model [**?**]. The nomenclature used here is adopted from a recent study by Sizov et al. [24] aimed at unifying the variations proposed for PLDA over the past decade.

A. Standard PLDA

The general idea of probabilistic linear discriminant analysis is to find a subspace in a given feature space that best resembles a subset of the original space's factors. In the case of speaker recognition, the subset of interest contains speaker specific components. The search for this subspace is based on a training dataset organized in a way that emphasizes differences between speakers as well as variations of each speaker across different recordings (aka sessions). The data organization is comprised of n_i observation i-vectors for speaker i from a set of development speakers, PLDA assumes the following linear factorization for each i-vector m_{ij}:

$$m_{ij} = \mathbf{V}y_i + \mathbf{U}x_{ij} + z_{ij}, \qquad j = 1, ..., n_i \tag{3}$$

where 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 representing eigenvoice and eigenchannel subspaces, respectively. Eigenvoice refers to the collection factor loadings (represented in $\mathbf{V}$) that construct the speaker-dependent subspace. Eigenchannel refers to the session-dependent subspace. In addition to the eigenchannel subspace a session-dependent and normally distributed slack variable, z_{ij}, is included to express session variabilities. In (3), z_{ij} takes a diagonal covariance matrix, $\mathcal{N}(0, \mathbf{\Sigma_d})$, [10], [9].

PLDA predicts model parameters, $(\mathbf{V}, \mathbf{U}, \mathbf{\Sigma_d})$, using the expectation-maximization (EM) algorithm [9]. After estimating subscpace components using background development data, trial i-vectors are then reduced to the same speaker-dependent subspace (indirectly) using PLDA and scored through a hypothesis testing procedure (see [9] for details). The hypothesis testing stage estimates the likelihood ratio of whether two trial i-vectors (train and test) belong to the same speaker, or if they belong to two different speakers.

B. Simplified PLDA

The second formulation reduces the complexity of (3) using the fact that session-dependent latent (x_{ij} in (3)) variables are not directly used in the scoring process. Therefore, as long as models are able to effectively estimate the eigenvoice subspace, channel related dimensionality (essentially all non-speaker dimensionality) is redundant. With this in mind, the second term in (3) is removed in simplified PLDA and all channel information is captured in the slack variable. The slack variable in this case is assumed to have a full covariance matrix.

$$m_{ij} = \mathbf{V}y_i + z^f_{ij}, \qquad j = 1, ..., n_i \tag{4}$$

The use of a full covariance matrix in simplified PLDA can be interpreted as combining the diagonal slack covariance in (3) with the eigenchannel subspace projections UU^T [24].

C. PLDA as an extension to LDA

PLDA can also be viewed as a probabilistic extension to LDA. The LDA based interpretation, also called the two-covariance model [**?**], [24], describes the i-vector space in terms of between- and within-speaker covariances. LDA models a feature space as a mixture of Gaussians, in which each mixture has the same covariance of Φ_w. The Gaussian mixtures represent the within class (i.e., speaker) variability, therefore Φ_w is refer to as the within-class covariance matrix. What LDA fails at is to provide a continuous (or in this context, stochastic) representation of each mixture's centroid. Therefore, centroids can be viewed as deterministic in LDA. PLDA provides a stochastic representation of class centroid by means of a between-class covariance matrix, Φ_b [25]. Using such an interpretation, PLDA can also be defined as two interdependent distributions;

1) the distribution of i-vectors in each class representing a certain speaker, which is a Gaussian with mean s, representing the average i-vectors corresponding to a speaker, and covariance Φ_w:

$$m \sim \mathcal{N}(s, \Phi_w), \tag{5}$$

2) the distribution of class centroids, also assumed Gaussian:

$$s \sim \mathcal{N}(s_G, \Phi_b), \tag{6}$$

where s_G is the global mean of all class centroids and Φ_b is the between-speaker covariance matrix.

Defining channel variability as a function of speaker variation helps PLDA model unseen speakers (i.e., speakers that are not present in the development set), which LDA is incapable of doing [25]. The interpretation in this section, provides a perspective which we will use in our proposed method (Sect. **??**) to include co-channel interference as a contributor to within-class variability. Equations (5) and (6) can be derived from (4) [24] or vice versa [25].

D. Baseline: mixed PLDA

The search for eigenvoice and eigenchannel subspaces involves a careful selection of development data. The idea in data preparation for PLDA is to provide sufficient channel diversity for each speaker to model within-speaker variations, while maintaining high speaker counts to be able to model between-speaker variability. Channel and speaker diversities introduced in the development data are directly translated into within- and between-speaker covariances, respectively. Said covariances are used to estimated PLDA parameters, (3) and (4) [**?**]. Such a data driven perspective towards compensating mismatch using PLDA has inspired a number of studies to tackle other types of variability through the same data selection procedure [26]. Thereby, instead of channel diversity, one could generate a development set with age diversity [**?**] or language diversity [26] (in the case of multi-lingual speakers). This leads to our first approach, which is to establish the amount of achievable performance gain in co-channel speaker recognition when background PLDA data contains co-channel speech samples. In this approach, recordings for each speaker consists of multiple co-channel samples. Note that it is important to maintain diversity in secondary (aka interfering) speakers, otherwise the PLDA model will train to both primary and secondary speakers in the co-channel sessions. Figure 7 is a diagram of how the PLDA data is arranged in this baseline approach, which we call "mixed PLDA". Mixed PLDA implicitly creates *co-channel awareness* in the background development data using co-channel mixtures.

Mixed PLDA describes the use of co-channel background

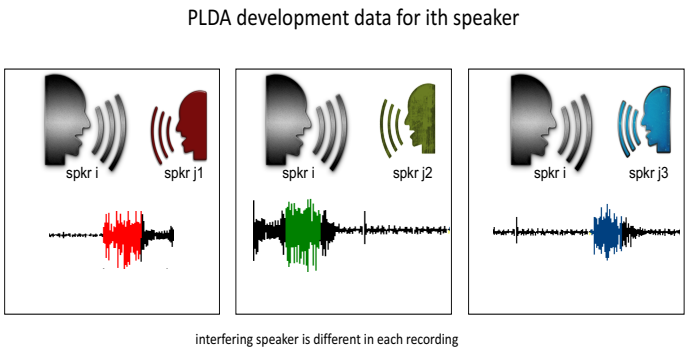

Fig. 7: *Creating development data for co-channel aware PLDA. the mixed PLDA approach uses co-channel data for each speaker in the background model. Recordings for the i^{th} speaker consists of co-channel sessions with different speakers.*

data in simplified and standard PDLA formulations of equations (4) and (3). This will be used as a baseline to provide fair comparison with our proposed PLDA formulation described in the next section, since (3) and (4) do not claim robustness with respect to co-channel data in their original forms.

IV. Proposed method: Co-channel Aware PLDA

This section describes our proposed approach, which is a modification to the PLDA formulation (Sect. III) that allows modeling speaker interference. Previously, in [27], we proposed a modified PLDA formulation to remove secondary speaker interference from the latent variable subspace. This study proposes a different method based on the two-covariance interpretation (described in Sect. III-C). The proposed method in this study, which we call *co-channel aware PLDA* (caPLDA), integrates the between-speaker covariance matrix into the within-class covariance to further emphasize the uncertainty caused by interfering speakers. Part of the reason for our departure from the co-channel PLDA formulation proposed in [27] was our inability to justify certain behavior in the model [**?**]. The current formulation (i.e., caPLDA), however, uses a sounder approach to combine co-channel interference with session variability.

caPLDA adopts the two-covariance interpretation of PLDA briefly described in (5) and (6). The i-vectors of a given class can be modeled as a normally distributed vector with a mean pertaining to the class to which it belongs and a covariance matrix representing within class variability, (5). Typically, within-class variability is meant to model channel variation across sessions. In the case of co-channel speech, in addition to channel variability, one must consider the variability caused by interfering speakers. In Sect. III-D, we attempted to capture speaker interference in the same manner channel variation is captured by PLDA. Although some improvement is observed, we expect that the original PLDA formulation, (**??**), is not capable of fully capturing speaker interference; partly due to the similarity between speaker-dependent latent variables and the cross session variations that exist in co-channel interference. We propose that in order to further improve performance, information can be shared between the speaker-dependent covariance matrix (between-speaker covariance) and the within-class covariance computed in the mixedPLDA method from Sect. III-D. For an i-vector with speaker A as the foreground speaker and some speaker X as secondary speaker, PLDA assumes this i-vector is generated from a normal distribution $\mathcal{N}m_A, \Phi$. While (III-C), assumes Φ is a unique within-class covariance matrix for all speakers in the i-vector space (i.e., Φ_w), we argue that an additional component is required to model within speaker variations in the case of co-channel i-vectors. The additional component contributing to within-class covariance is of the same nature of the between-class covariance. Therefore, one can assume that Φ is a function of both Φ_w and Φ_b (as defined in Sect. III-C), $\mathcal{F}(\Phi_w, \Phi_b)$. Our suggested structure for $\mathcal{F}(.,.)$ is a linear combination of the two covariance matrices:

$$\Phi = \mathcal{F}(\Phi_w, \Phi_b) = \Phi_w + \alpha\Phi_b \tag{7}$$

where α is a function of the signal-to-interference ratio between the foreground and background speaker. For clean (non co-channel) i-vectors, α should be 0, while α increases as SIR

drops. Note that the typical scale of the Φ_b is much large than Φ_w [28], therefore the range of $\alpha \in (0, \alpha_{max})$ is chosen such that $\alpha_{max} << 1$.

V. Conclusion

The conclusion goes here.

References

[1] 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.

[2] R. E. Yantorno, "Cochannel speech study," Electrical and Computer Engineering Department Temple University, Tech. Rep., September 1999.

[3] 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.

[4] 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.

[5] N. Shokouhi, A. Sathyanarayana, S. Sadjadi, and J. H. L. Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in *Proc. IEEE ICASSP*, Vancouver, BC, May 2013.

[6] 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.

[7] 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.

[8] O. Cetin and E. Shriberg, "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.

[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] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] NIST, "The NIST year 2004 speaker recognition evaluation plan," 2008. [Online]. Available: http://www.nist.gov

[18] ——, "The NIST year 2005 speaker recognition evaluation plan," 2008. [Online]. Available: http://www.nist.gov

[19] ——, "The NIST year 2006 speaker recognition evaluation plan," 2008. [Online]. Available: http://www.nist.gov

[20] 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.

[21] NIST, "The NIST year 2008 speaker recognition evaluation plan," 2008. [Online]. Available: http://www.nist.gov

[22] S. O. Sadjadi and J. H. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[23] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Bořil, and J. H. Hansen, "Crss systems for 2012 nist speaker recognition evaluation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6783–6787.

[24] A. Sizov, K. A. Lee, and T. Kinnunen, "Unifying probabilistic linear

discriminant analysis variants in biometric authentication," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, 2014, pp. 464–475.

[25] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.

[26] A. Misra and J. H. Hansen, "Spoken language mismatch in speaker verification: An investigation with nist-sre and crss bi-ling corpora," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 372–377.

[27] N. Shokouhi and J. H. Hansen, "Probabilistic linear discriminant analysis for robust speaker identification in co-channel speech," in *Sixteenth Annual Conference of the International Speech Communication Association, INTERSPEECH*, Dresden, Germany, 2015.

[28] O. Glembek, J. Ma, P. Matjka, B. Zhang, O. Plchot, L. Brget, and S. Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4032–4036.