MULTI-SPEAKER CONVERSATIONS, CROSS-TALK, AND DIARIZATION FOR SPEAKER RECOGNITION

Gregory Sell and Alan McCree

Human Language Technology Center of Excellence
The Johns Hopkins University, Baltimore, MD 21218, USA

ABSTRACT

I-vector training and extraction assume that a speech file is spoken by a single speaker. This work considers the effects of violating that assumption with the presence of cross-talk or multi-speaker conversations. First, it is demonstrated that these problematic speech files can be detected using the i-vector representation itself. The impact of these violations of the single-speaker assumption are then explored along with strategies to mitigate it. It is shown that, even in predominantly clean data, the removal of cross-talk can provide modest gains, but that T matrix and PLDA training are largely robust to these types of noise. It is also shown that detection in front of diarization is a reasonable strategy in the presence of data with an unknown proportion of multi-speaker conversations. Finally, in the course of this work, evidence is found that cross-talk detection and multi-speaker detection may in fact be different tasks that require separately trained detectors.

Index Terms— speaker diarization, speaker recognition, i-vectors

1. INTRODUCTION

Speech technology researchers often make assumptions about their audio in the process of exploring new methods and ideas. Sometimes, these assumptions are even explicitly capitalized on to improve performance, such as speaker adaptation in automatic speech recognition (ASR) assuming that the speech originates from a single speaker. But, these assumptions may not always hold true in noisy data, and so front-end mitigation can be required. For the single-speaker assumption, the front-end mitigation is speaker diarization, which groups speech into segments spoken by the same person.

And speaker adaptation is not the only technology that expects speech from a single speaker. I-vectors, the audio representation widely used for speaker and language recognition as well as speaker adaptation in recent ASR algorithms, assume in both training and test-time embedding that the speech in a given file all comes from the same speaker. So, the presence of multiple speakers will violate this assumption and likely lessen the effectiveness of the representation. This effect has clearly been shown in the past [1, 2] for multi-speaker test conversations. In [2], the speaker recognition error rate increased by over 50% (relative) without diarization. Interest in this topic was recently renewed with the 2016 Speakers in the Wild Challenge [3], where it was shown again that diarization can improve speaker recognition rates in uncontrolled audio [4, 5].

The work that follows more expansively explores the effect of violating the single-speaker assumption, considering both cross-talk and multi-speaker conversations, and the full i-vector speaker recognition pipeline is examined, including training and testing.

The discussion will first focus on the detection of the problematic files. Ideally, these conversations could be easily found by diarizing all files and assuming that the clean, single-speaker files will simply be found to have only one speaker. As later results will show, this assumption is not necessarily valid, and unnecessary diarization can potentially degrade performance. Furthermore, from an efficiency point of view, state-of-the-art i-vector segment diarization is also a potentially expensive process, requiring the extraction and clustering of hundreds of i-vectors for each file. A simple classifier front-end that efficiently identifies the need (or absence of need) for diarization would allow for more efficient data processing.

An additional aspect of this work is the consideration of cross-talk in the audio. Cross-talk in conversational telephone speech refers to the undesired leakage of one side of the call into the audio of the other side. As a result, cross-talk is a potentially different condition from a multi-speaker conversation, because, in cross-talk, the interfering speech is often significantly distorted and degraded, unlike in a true multi-speaker conversation where all speakers are processed through similar channels.

Several related tasks have been explored in the past, but, in most cases, that work was looking to identify usable frames in the presence of overlapping speech, rather than identify entire conversations that include multiple speakers. These detectors were intended as a front-end for various technologies such as speaker identification [6] or speaker diarization [7], and there is also evidence that these types of detections would benefit speech recognition as well [8]. In past work, experiments have tested detectors for overlapping speech that utilize a wide range of features, including spectral autocorrelations [9], kurtosis [10] gammatone subband frequency modulation features [11], and multi-pitch tracking [6]. However, in this work, we aim to use features that are already computed in the i-vector pipeline, in order to minimize the cost of the detection process.

The task of speaker counting is actually a closer match to the current application, where the number of speakers in a conversation are estimated. In our current scenario, the counter would be reduced to a binary decision, separating conversations into single-speaker and multi-speaker, but the overall goal is more closely matched here than with overlap detection. In past work, speaker counting has typically been attempted with full speaker diarization systems [12, 13, 14], while we seek a more efficient solution.

The work that follows will first explore the tasks of cross-talk detection and multi-speaker conversation detection. These experiments will be followed by several that consider the effect these detections can have on downstream i-vector tasks, either with filtering of training lists or diarization of test files. Through the course of these experiments, it will be shown that cross-talk detection can improve performance even in clean data sets, while multi-speaker diarization can significantly reduce the diarization computation needed in the face of multi-speaker conversations. Additionally, results will

Feature	EER (Raw)	EER (Length Norm)
Zero-Order Stats	8.38	6.45
First-Order Stats	4.91	4.97
Second-Order Stats	5.48	5.79
I-Vectors	9.98	**4.80**

Table 1. Performance for SVM classification of the sufficient statistics and resulting i-vector for cross-talk detection in terms of equal error rate (EER). All representations perform reasonably well, though length-normalized acoustic i-vectors yield the lowest errors.

suggest that cross-talk detection and multi-speaker conversation detection may be more different tasks than intuition would suggest.

2. CROSS-TALK DETECTION WITH I-VECTORS

If detection of cross-talk and multi-speaker conversations is desired in an i-vector pipeline, it is sensible and efficient to utilize the analytics already processed in the task to identify the problematic segments. In the case of i-vectors, these representations include the acoustic features, the sufficient statistics (zero-, first-, and second-order) computed from the UBM, and the extracted i-vectors themselves. Effective detection of cross-talk using any of these features would allow for adding this process to the i-vector pipeline at very little computational cost.

2.1. Automatic Cross-Talk Labeling in Switchboard I

In order to measure the ability of each of these feature types to detect cross-talk, an evaluation on Switchboard I was built. Cross-talk is anecdotally known to be present in many conversation sides in the Switchboard I corpus, but labels identifying the distorted sides are not automatically provided. So, to estimate these labels, a simple spectral comparison across conversation sides was used. First, turn-taking was established with word-level forced alignments from the transcripts. Then, the presence of the speech from the active side in the conversation was estimated in the inactive side with a scalar-weighted least-squares optimization.

$$\beta = \min_{\alpha} \sum_{f,t \in T_A} (\alpha X_A[f,t] - X_B[f,t])^2$$

where f and t are frequency and time bins, respectively, and T_A is the set of time frames where side A is the active speaker. Cross-talk can then be detected by identifying conversation sides with a large multiplicative coefficient α, which estimates the strength of the active speech in the inactive conversation side, and a small ratio of the solution β to the original energy in the inactive spectrum X_B, which estimates the remaining energy in the inactive side after removing the speech. Through some informal listening experiments, it was determined that $\alpha > 0.01$ and $\beta < 0.95$ yielded reasonable estimates of cross-talk labels. With this process, 2,975 of the 4,870 sides in Switchboard I that were tested were labeled with cross-talk. The labels extracted with this process will undoubtedly have noise, but they allow for some instructive exploration of detection features. The detection systems will then be later tested within the context of downstream applications with more manicured evaluation labels.

Train	X-Talk (Swbd)	Multi (Fisher)	Multi (SRE)
X-Talk (Swbd)	4.80	28.12	27.12
Multi (Fisher)	17.58	11.84	11.85
Multi (SRE)	17.15	13.33	9.83

Table 2. EER rates for detecting cross-talk and multi-speaker conversations. All experiments use length-normalized i-vectors with SVMs. There is a clear mismatch between the detectors and the tasks, with each performing much better at its own task.

2.2. Cross-Talk Detection Features

The labels estimated for cross-talk in Switchboard I can then be used to compare feature representations for detection. Specifically, we will consider the sufficient statistics computed in the i-vector extraction process along with the extracted i-vectors themselves. The i-vector system used here and in all following experiments trained its universal background model (UBM) and T matrix on Fisher English data. The UBM consists of a 2048-component GMM with diagonal covariances, and the T matrix subsequently projects the representation to 600 dimensions. Probabilistic Linear Discriminant Analysis (PLDA) trained on data from the NIST SRE 04, 05, 06, and 08 data further reduces the dimensionality to 200 dimensions.

The results for the cross-talk detection with zero-order statistics (counts), first-order statistics (means), second-order statistics (covariances), and i-vectors can be seen in Table 1. In all cases, the features were classified with support vector machines (with output calibrated into probabilities with Platt scaling [15]), and each feature's results are shown with and without length normalization [16]. The Switchboard data was broken into 5 folds, and evaluation was performed with cross-fold validation.

The results in Table 1 show that the cross-talk labels can indeed be learned well with the existing analytics from the i-vector process. Of the statistics, the first-order are most effective, followed closely by the second-order, while the zero-order statistics give the highest error rates. However, the i-vectors themselves are most effective once length-normalized, which is not surprising given the success of the first-order statistics and the close relationship between those two representations.

Based on their performance, the experiments to follow will use the length-normalized i-vector representations for cross-talk and multi-speaker detection. For completeness, several scalar analytics derived from the statistics, such as frame counts and entropy of the zero-order statistics, were also explored, but none achieved an error rate below 22%, and the fusion of all scalar features still only yielded a 17.4% error rate, all substantially worse performances than the length-normalized i-vectors.

3. MULTI-SPEAKER CONVERSATION DETECTION WITH I-VECTORS

Detecting speech with more than one speaker is theoretically very similar to detecting cross-talk, but, as discussed above, they may indeed be different tasks. To explore this relationship, a separate set of training corpora were assembled to detect multiple speakers from Fisher English and NIST SRE '04, '05, '06, and '08. These data sets were selected because they are expected to be less corrupted by cross-talk than Switchboard, and we desire data without cross-talk in order to study the differences between the two tasks.

For each corpus (Fisher or SRE), the two sides of a call were

T Matrix Training List	EER	DCF
Fisher	2.57	**0.230**
Fisher (no x-talk)	2.74	0.241
Fisher (no multi-spkr)	2.57	0.235
Fisher+Swbd	2.59	0.243
Fisher+Swbd (no x-talk)	**2.46**	0.240
Fisher+Swbd (no multi-spkr)	2.61	0.246
Fisher+SummedSRE	2.51	0.230
Fisher+SummedSRE (no x-talk)	2.51	0.240
Fisher+SummedSRE (no multi-spkr)	2.55	0.236

Table 3. EER and DCF on SRE10 for various training lists. Only small effects are seen from the removal of multi-speaker conversations or cross-talk, even after artificial addition of those corruptions.

summed for a randomly selected subset. This results in roughly 67% of files with one speaker and the remaining 33% with two speakers.

Detection results are shown in Table 2, using the assembled corpora from Fisher and SRE for multi-speaker detection and the Switchboard labels for cross-talk. Note that, as with the previous cross-talk experiments, systems trained and tested on the same data used cross-fold validation with five folds.

We see in Table 2 that multi-speaker detection is possible with the i-vector features, although the error rates are roughly double that of the cross-talk task. Interestingly, the fact that i-vectors can be used for these detections demonstrates that the presence of cross-talk or multiple speakers impacts the i-vector representation.

These results also suggest that the system trained for cross-talk detection struggles in identifying the summed conversations, and the multi-speaker systems similarly struggle on cross-talk detection. One explanation of these results is that cross-talk detection and multi-speaker conversation detection are actually significantly different tasks. It is also possible that these discrepancies are due to cross-corpus domain mismatch rather than variations in the task, but later results in speaker recognition will indicate that the cross-talk detector does appear to detect problematic audio across corpora. Furthermore, the two multi-speaker detectors behave very similarly despite being trained on different corpora.

4. CROSS-TALK AND MULTI-SPEAKER DETECTION FOR I-VECTOR APPLICATIONS

The above sections explored the efficacy of i-vector representations for cross-talk and multi-speaker conversation detection. While the results suggest that the architecture does indeed have some value for the task, simply detecting these conditions is not necessarily a useful outcome. Instead, we next explore the impact these detections can have in the context of downstream speaker recognition.

4.1. Total Variability Subspace Training

First we consider the training of the T matrix, during which it is assumed that each audio recording has only one speaker and one channel. The presence of cross-talk or multiple speakers would violate this assumption and potentially hinder the training of the parameters.

To explore this hypothesis, we tested an acoustic i-vector speaker recognition system on the NIST SRE10 evaluation, measuring equal error rate (EER) and the minimum detection cost function (DCF) to compare performances. Unless otherwise stated, all systems used the Fisher English corpus to train the UBM, and the NIST SRE '04, '05, '06, and '08 data to train the PLDA. Each system was then varied according to the training data for the total variability subspace (T matrix).

The initial system was trained on the entire Fisher English corpus, and then a second and third system were trained with the Fisher English list filtered down to exclude any files determined to have cross-talk by the Switchboard-trained detector or any files determined to have multiple speakers by the Fisher-trained detector, respectively. These results are shown in the first several entries of Table 3. Neither reduction to the list improves the performance, and, in fact, the cross-talk removal negatively affects the EER while both hurt the DCF.

However, it is possible that these underwhelming results are related to the cleanliness of the Fisher English data. It is hard to improve performance by removing speech with cross-talk or multiple speakers if there aren't many examples to start with. So, to ensure the presence of distortions, a second set of lists appended the Switchboard data, which is known to have cross-talk, or the artificially summed SRE data to the Fisher English training list, as well as appending the subset of Switchboard believed to be clean of cross-talk or multi-speaker conversations according to the i-vector-based detector to the Fisher English list.

The results for these systems can also be seen in Table 3. While adding Switchboard to the training list actually slightly hurts in both metrics versus the Fisher-only list, the removal of cross-talk from this list yields improvements. While the DCF performance is only a modest improvement on the full Fisher+Switchboard list, the EER improves by a larger margin and even slightly outperforms the baseline. So, while adding all of Switchboard to the T matrix training hurts overall performance, adding only the screened subset can actually improve it. Multi-speaker detection offers no help for this list. Adding summed SRE data to Fisher actually gives a small improvement on its own, but neither detection task improves performance.

Similar tests were run to explore PLDA training, but little effect was seen from altering the training lists.

4.2. Cleaning Test Segments

While the effects of cross-talk and multi-speaker conversations in subspace training has not previously been thoroughly explored, previous experiments have demonstrated the negative effect that they can have on speaker recognition trials [1, 2, 4, 5]. In past work, the natural solution to this problem has been to diarize the multi-speaker conversation, compare each resulting speaker to the speaker model, and keep the maximum score across all diarized speakers to represent the log likelihood ratio of the entire file. We will employ the same strategy here when diarization is used.

However, past experiments have only considered the case where all conversations have multiple speakers, and it is known in advance that all conversations require diarization. We instead consider the case where an unknown subset of conversations have multiple speakers, and so simply running diarization on all conversations may not yield optimal results. Furthermore, diarization is a costly process, and running it unnecessarily may be undesirable from an efficiency standpoint in addition to potential performance ramifications.

In our experiment, we consider several scenarios:

- Original - the original SRE10 test list with no summed conversations
- Rare - 5% of test list summed to multi-speaker
- Half - 50% of test list summed to multi-speaker
- Full - 100% of test list summed to multi-speaker

	Original			Rare			Half			Full		
	EER	DCF	% Diar.	EER	DCF	% Diar.	EER	DCF	% Diar.	EER	DCF	% Diar.
No Diarization	2.57	0.230	0.0	4.06	0.275	0.0	10.22	0.504	0.0	14.90	0.745	0.0
Diarize X-Talk	**2.47**	**0.229**	5.0	3.65	0.266	6.1	9.12	0.459	12.8	13.40	0.658	21.0
Diarize Multi-Spkr	2.69	0.240	13.6	3.12	0.261	17.6	5.21	0.352	47.4	7.30	0.444	80.4
Diarize Both	2.62	0.238	17.1	3.05	**0.259**	21.3	5.06	0.350	50.7	6.98	0.440	82.4
Diarize All	2.74	0.259	100.0	**3.01**	0.277	100.0	**4.25**	**0.338**	100.0	**4.89**	**0.395**	100.0

Table 4. SRE10 results in terms of EER and DCF for several test conditions. Several strategies for diarization are also shown along with the percentage of the data that was diarized for that particular strategy. In the presence of little to no multi-speaker conversations, performance can be improved over the baseline with significantly reduced computation costs (compared to diarizing all test files).

In each of these test cases, we will also explore several strategies for speaker recognition using diarization. Note that only test cuts were made available for diarization. Enrollment files were not summed in any of the scenarios, and were not diarized in any of the strategies.

- No diarization
- Diarize only files flagged by cross-talk detection
- Diarize only files flagged by multi-speaker detection
- Diarize files flagged by either detector
- Diarize all test files.

For all diarization experiments, an acoustic i-vector system was used that extracts i-vectors for each second of speech and clusters them with agglomerative hierarchical clustering, with the system trained with NIST SRE data as described in [17]. Speaker priors restricting the decision to one or two speakers were incorporated as well[14].

Results for all these experiments can be seen in Table 4. In the case of the original SRE10 list, there are two interesting observations. First, the diarization of files detected to have cross-talk yields small improvements in both metrics, while all other diarization strategies degrade performance. The second observation is that diarizing all files hurts performance, but not as significantly as one might expect, especially when one considers that the diarization algorithm determined there was only one speaker in less than 1% of the single speaker files. This lack of degradation is likely in part due to the long durations of these files (often several minutes), and so dividing the speech into several groups will still leave each with sufficient data for an effective speaker recognition decision. For data with less speech, the consequences could be more severe.

Once multi-speaker conversations are introduced into the test condition, even in the case where only 5% are summed, running no diarization becomes the worst performing system. And, when 50% or 100% of the conversations are summed, omitting diarization becomes nearly catastrophic, increasing errors by up to 6 times. Alternatively, diarizing all conversations quickly becomes the optimal strategy, giving best or nearly best performance in all conditions that include any summed conversations at all. However, there is a clear trade-off that can be seen when considering the amount of the test set that is diarized. For the rare condition, for example, diarizing only those cuts with cross-talk or multiple speakers detected yields very similar performance to diarizing all files, but requires only processing $\sim$20% of the test set. So, in this case, a large reduction in computation is achieved with essentially no impact on performance. As the prevalence of summed conversations increases, the trade-off tips more in favor of diarizing everything, as the performance gap widens while the computational savings decrease. In practice, if multi-speaker conversations are expected to be dominant, then diarizing all files appears to be the best solution. However, if multi-speaker conversations are expected to be less common or if there are computational concerns, then diarizing only detected files is likely the better approach.

It is also an interesting observation that the baseline of no diarization is not the best performing system in any condition. This result suggests that diarization's role in i-vector technologies should perhaps be reconsidered and potentially increased.

Table 4 taken in aggregate also provides some insight into a previous observation, which is that the cross-talk and multi-speaker systems struggle with each other's tasks in Table 2. However, in that context, it was difficult to determine whether that outcome was related to mismatch between train and test data, or whether it was related to fundamental differences in the two seemingly related tasks. The results in Table 4 provide additional evidence that the tasks themselves are different. In the original condition, the cross-talk detector yields performance improvements despite being trained on the less-matched data. However, in the presence of multi-speaker conversations, the cross-talk diarization barely outperforms the baseline of no diarization at all, suggesting that the detector, while effective for cross-talk in this task, is ineffective for detecting summed conversations. The multi-speaker detector, on the other hand, degrades performance in the absence of multi-speaker conversations, but makes much greater improvements than the cross-talk system once summing begins. So, within this task, each detector performs better in the task it was trained for, despite all experiments being run on the same data. This is not conclusive, but it is strong evidence that cross-talk detection and multi-speaker detection are indeed different tasks.

5. CONCLUSION

This work explored the task of detecting speech files with cross-talk or multiple speakers. It was demonstrated that these tasks can be effectively performed with i-vectors, and that proper use of the detections can improve speaker recognition in a variety of conditions. At test time, in particular, it was shown that diarization with a detector front-end improved performance in all tested conditions, including the SRE10 evaluation data itself. It was also shown that modest improvements could be achieved by removing cross-talk from T matrix training lists, though overall this stage of the pipeline as well as PLDA training appeared to be largely robust to the presence of cross-talk or multi-speaker conversations, even when artificially added. The experiments also yielded results suggesting that cross-talk detection and multi-speaker conversation detection may be different enough tasks that separate detectors are required for each. Taken in aggregate, these results provide a compelling argument for the importance of mitigating the presence of audio data that violates the single-speaker assumption. The results also show that a combination of detection and diarization can help reduce the effects of these unexpected speakers in i-vector applications.

6. REFERENCES

[1] Alvin F. Martin and Mark A. Przybocki, "Speaker Recognition in a Multi-Speaker Environment," in *Proceedings of Interspeech*, 2001.

[2] Patrick Kenny, Douglas Reynolds, and Fabio Castaldo, "Diarization of Telephone Conversations using Factor Analysis," *IEEE Journal of Special Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–70, December 2010.

[3] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson, "The 2016 Speakers in the Wild Speaker Recognition Evaluation," in *Proceedings of Interspeech*, 2016.

[4] Ondřej Novotný, Pavel Matějka, Oldřich Plchot, Ondřej Glembek, Lukáš Burget, and Jan "Honza" Černocký, "Analysis of Speaker Recognition Systems in Realistic Scenarios of the SITW 2016 Challenge," in *Proceedings of Interspeech*, 2016.

[5] Yi Liu, Yao Tian, Liang He, and Jia Liu, "Investigating Various Diarization Algorithms for Speaker in the Wild (SITW) Speaker Recognition Challenge," in *Proceedings of Interspeech*, 2016.

[6] Yang Shao and DeLiang Wang, "Co-Channel Speaker Identification Using Usable Speech Extraction Based on Multi-Pitch Tracking," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[7] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Acoustics, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–70, February 2012.

[8] Elizabeth Shriberg, Andreas Stolcke, and Don Baron, "Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation," in *Proceedings of Interspeech*, 2001.

[9] Katsuri Rangan Krishnamachari, Robert E. Yantorno, Daniel S. Benincasa, and Stanley J. Wenndt, "Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments Under Co-Channel Conditions," in *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, 2000.

[10] Kasturi Rangan Krishnamachari, Robert E. Yantorno, and Jereme M. Lovekin, "Use of Local Kurtosis Measure for Spotting Usable Speech Segments in Co-Channel Speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001.

[11] Navid Shokouhi, Seyed Omid Sadjadi, and John H. L. Hansen, "Co-channel Speech Detection via Spectral Analysis of Frequency Modulated Sub-bands," in *Proceedings of Interspeech*, 2014.

[12] Ali Ziaei, Abhijeet Sangwan, and John H. L. Hansen, "Prof-Life-Log: Personal Interaction Analysis for Naturalistic Audio Streams," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[13] Mohammed Senoussaoui, Patrick Kenny, Themos Stafylakis, and Pierre Dumouchel, "A Study of the Cosine Distance-Based Mean Shift for Telephone Speech Diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 217–27, January 2014.

[14] Gregory Sell, Alan McCree, and Daniel Garcia-Romero, "Priors for Speaker Counting and Diarization with AHC," in *Proceedings of Interspeech*, 2016.

[15] John C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

[16] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of Interspeech*, 2011.

[17] Gregory Sell and Daniel Garcia-Romero, "Speaker Diarization with PLDA I-Vector Scoring and Unsupervised Calibration," in *Proceedings of the IEEE Spoken Language Technology Workshop*, 2014.