# Investigating Various Diarization Algorithms for Speaker in the Wild (SITW) Speaker Recognition Challenge

*Yi Liu, Yao Tian, Liang He, Jia Liu*

Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
{liu-yi15, tianyao11}@mails.tsinghua.edu.cn, {heliang, liuj}@tsinghua.edu.cn,

## Abstract

Collecting training data for real-world text-independent speaker recognition is challenging. In practice, utterances for a specific speaker are often mixed with many other acoustic signals. To guarantee the recognition performance, the segments spoken by target speakers should be precisely picked out. An automatic detection could be developed to reduce the cost of expensive human hand-made annotations. One way to achieve this goal is by using speaker diarization as a pre-processing step in the speaker enrollment phase. To this end, three speaker diarization algorithms based on Bayesian information criterion (BIC), agglomerative information bottleneck (aIB) and i-vector are investigated in this paper. The corresponding impacts on the results of speaker recognition system are also studied. Experiments conducted on Speaker in the Wild (SITW) Speaker Recognition Challenge (SRC) 2016 showed that the utilization of a proper speaker diarization improves the overall performance. Some more efforts are made to combine these methods together as well.

**Index Terms**: speaker recognition, speaker diarization, speaker in the wild

## 1. Introduction

Speech technologies have been developed rapidly during last decades. The applications of speaker recognition now attract much more attention. By virtue of i-vector framework and deep learning, the accuracy of speaker recognition significantly improved [1]. Even though the performance of state-of-the-art system in controlled condition is satisfying, it suffers a deterioration in real-world scenario. The background noise [2], reverberation [3], channel mismatch [4] and compression artifacts will all influence the quality of recordings. Moreover, the flexible duration [5], mixture of genders, intra-speaker variability due to physiological status [6], also exhibit negative impacts on the practical performance.

Besides, collecting enough audio to train robust speaker models is another problem. Reliable authentication needs minutes of target speaker voice to enroll a model. In practice, the captured audio often contains multiple speakers talking in free-style, like telephone conversations and interviews. Separating the speech for target speakers is challenging. It is labor-intensive to annotate multi-speaker audio by human and wastes plenty of time. Some automatic methods should be developed to locate excerpts spoken by target speakers.

Some speaker diarization algorithms have been involved to solve the problem. Speaker diarization aims to answer "who spoke when" and is able to separate different speakers in one utterance. Dozens of researches have been done in this field [7, 8]. Unfortunately, these works mainly focused on the diarization error rate (DER) which only reports the correctness of diarization [9, 10, 11]. However, because we only care about target speakers, DER shows no necessary connection with the performance of the back-end speaker recognition. For example, a method that selects audio excerpts with higher purity of target speakers, would probably perform better than the one which achieves lower DER while mixes more non-target segments in the enrollment data.

In previous NIST speaker recognition evaluation (SRE), summed-channel telephone conversations are included as optional training and test conditions. Some speaker recognition experiments have already done on these conditions. But we noticed that the published papers did not compare different speaker diarization algorithms in their proposed systems [12, 13, 14]. Also, due to the limited channel conditions in NIST SRE, the combination of speaker diarization and recognition misses tests under more wild circumstances.

Recently, SRI released the speaker in the wild (SITW) speaker recognition challenge (SRC) database [15]. The recordings in this database are real-world collected and contain both single- and multiple-speaker data. This database gives us a great opportunity to investigate various diarization algorithms in speaker recognition systems. In this paper, we first introduce our i-vector system developed for SITW SRC. Then, three important diarization algorithms are implemented. Their impacts on the final performance of speaker recognition are evaluated under different conditions. We hope this paper could give the society some empirical understandings about the effects of different diarization methods in speaker recognition systems.

The organization of this paper is as follows. The i-vector speaker recognition system we developed is briefly introduced in Section 2. Section 3 describes three speaker diarization algorithms. Our development data, parameter setup and experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

It should be pointed out that the results in this paper were achieved after the deadline of the SITW challenge.

## 2. System framework

A standard i-vector system was built during the post-evaluation period. The system flowchart is shown in Figure 1. Each part is discussed below.

### 2.1. Feature extraction

The pre-processing stage of our system consisted of voice active detection (VAD) and feature extraction. Non-speech frames are gated using a sub-band entropy VAD algorithm. Static percep-

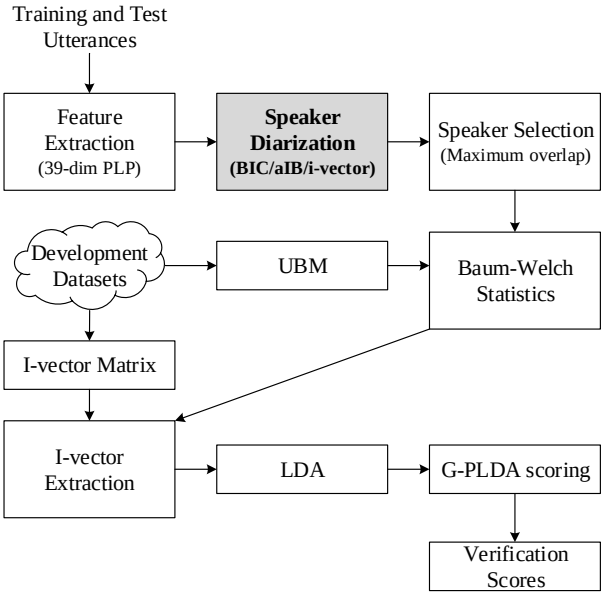


Figure 1: *The flowchart of our i-vector speaker recognition system. In this paper, we focus on the diarization module, which is emphasized by a shaded block.*

tual linear prediction cepstral coefficients (PLP) with log energy, appended with delta and delta-delta derivations, are extracted as features. Feature warping is applied at last [16].

### 2.2. Speaker diarization and selection

For audio containing multiple speakers, three diarization methods were investigated. The algorithms are described in detailed in the next Section. We selected the speaker clusters with maximum overlap with the given annotations. Non-target segments were discarded.

### 2.3. Modeling

A gender-independent universal background model (UBM) was trained and then used to collect Baum-Welch statistics for i-vector modeling. The UBM and i-vector $\boldsymbol{T}$ matrix were both trained using EM algorithm.

### 2.4. Classifier

After i-vectors were extracted, linear discriminant analysis (LDA) [17] and Gaussian probabilistic linear discriminant analysis (G-PLDA) scoring [18] were applied. The LDA projection matrix and G-PLDA were trained using data from background speakers. The verification scores were the log-likelihood ratio of the target and non-target hypotheses.

## 3. Speaker diarization algorithms

In specific conditions of SITW SRC, training utterances contain multiple speakers and noise. If irrelevant speech mixed in, the quality of models would certainly decrease. In this work, three important speaker diarization methods are studied to locate speakers of interest precisely. The algorithms are presented below.

**Algorithm 1** BIC-based diarization

1: **Segmentation**
1a. Set a pair of adjacent windows in the beginning of the utterance.
1b. Calculate $\Delta$BIC in the current position.
1c. Slide the adjacent windows on the utterance and compute $\Delta$BIC repeatedly.
1d. Find local maxima along $\Delta$BIC. These points indicate the potential speaker change points.
1e. Split audio in these positions.
2: **Clustering**
2a. Treat each segment as a cluster. Calculate $\Delta$BIC between any possible combinations of current clusters.
2b. Merge segments with the lowest distances.
2c. Update clusters and re-compute the distance matrix.
2d. Repeat 2a-2c until $K$ clusters left. $K$ is the number of speakers in hypothesis.
3: **Viterbi decoding**
3a. Use $K$ clusters to generate speakers' Gaussian mixture models (GMM) (with relevance MAP adaptation).
3b. Calculate log-likelihoods for all speech frames given different GMMs.
3c. Decode using Viterbi to find the best path.

### 3.1. Bayes information criterion-based algorithm

Tranter and Reynolds introduced a general diarization framework in [7]. This framework could be implemented by three conventional steps: change point detection based on Bayes information criterion (BIC), BIC-based clustering and Viterbi decoding.

Given two sets of observations, whether they are generated a single model can be measured by BIC formulation

$$\begin{aligned}\Delta\text{BIC} =& \frac{1}{2}\left[N_z \log|S_z| - N_x \log|S_x| - N_y \log|S_y|\right] \\ & - \alpha P\end{aligned} \tag{1}$$

where $x$ and $y$ denote two Gaussian models estimated from the distinct sets, while $z$ denotes a single Gaussian which models all data together, $N$ and $S$ are the number of samples and the corresponding covariance matrices, respectively. In addition, $P = [d(d+3)/4] \cdot \log N_z$ is the penalty term and $d$ is the dimension of the observations. The lower $\Delta$BIC is, the more likely the two sets are generated from one single model.

Equation 1 can be used to measure the similarity between two speech segments, thus motivates the BIC-based diarization. The steps are summarized in Algorithm 1.

### 3.2. Agglomerative information bottleneck method

Information bottleneck (IB) is a powerful clustering framework in information theory and has been applied in many tasks [19]. IB is first introduced to speaker diarization by Vijayasenan [9]. Given a dataset $X$, and the *relevance variables* $Y$, we try to find a compact and informative representation $C$ that maximize mutual information $I(Y,C)$ under a constraint on $I(X,C)$. The representation $C$ is like a bottleneck that the information contains in $X$ about $Y$ passed through.

When applying IB into speaker diarization, speech is first split into segments with equal length. The features in segment $i$, denoted as $\boldsymbol{x}^{(i)}$, are variable $X$. The relevance variables $Y$ are defined as the components of a GMM $\Lambda = \{w_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^{M}$

**Algorithm 2** aIB diarization

1: **Preparation**
1a. Split the speech features equally, hence get $M$ segments.
1b. GMM estimation based on the segmentation.
1c. Compute $P(y_j|\boldsymbol{x}^{(i)})$ for each segment $i$ and component $j$.
2: **Clustering**
Treat $P(y_j|\boldsymbol{x}^{(i)})$ as features and apply aIB clustering algorithm.
3: **Viterbi realignment**
Viterbi decoding using KL-divergence metric with minimum duration constraint [20].

estimated on the all segments. The mixture of the GMM is $M$, where $M$ equals to the total number of segments.

Apparently, the probability $P(y_j|\boldsymbol{x}^{(i)})$ that the $N$-frame segment $\boldsymbol{x}^{(i)}$ belonging to component $y_j$ is expressed by

$$P(y_j|\boldsymbol{x}_n^{(i)}) = \frac{w_j \mathcal{N}(\boldsymbol{x}_n^{(i)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{M} w_l \mathcal{N}(\boldsymbol{x}_n^{(i)}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \tag{2}$$

$$P(y_j|\boldsymbol{x}^{(i)}) = \frac{1}{N}\sum_{n=1}^{N} P(y_j|\boldsymbol{x}_n^{(i)}) \tag{3}$$

where $\boldsymbol{x}_n^{(i)}$ is the $n$-th sample of segment $i$. Given $X$, $Y$ and $P(Y|X)$, agglomerative IB (aIB) could be applied according to [9]. Our implementation is shown as Algorithm 2.

### 3.3. Simplified i-vector-based diarization

The i-vector-based diarization is inspired by the fact that speaker information is effectively conveyed by low-dimension vectors. Some works have shown that one can achieve better DER using this method. Principal component analysis (PCA) is also exploited to improve the performance [21].

We denote the i-vectors of all segments as $\{\boldsymbol{w}_i\}_{i=1}^{M}$. PCA is applied on these i-vectors and let $\lambda_i$ and $\boldsymbol{u}_i$ denote the $i$-th eigen-value and the corresponding eigen-vector in decreasing order. The PCA-projected vector is denoted as $\boldsymbol{w}_i'$, and further transformed by

$$\hat{\boldsymbol{w}}_i = \boldsymbol{\Lambda}^{1/2}\boldsymbol{w}_i' \tag{4}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix containing eigen-values. Then reduce the dimension of $\hat{\boldsymbol{w}}_i$ to $d_0$, which is dynamically decided by

$$d_0 = \arg\min_d \frac{\sum_{i=1}^{d}\lambda_i}{\sum_{j=1}^{D}\lambda_j} \geq t_0 \tag{5}$$

where $D$ is the dimension of $\hat{\boldsymbol{w}}_i$ and $t_0$ is a pre-determined threshold. K-means is usually applied to these new vectors.

In our experiment, we simplified this i-vector-based diarization, and ignored the Viterbi realignment and second pass refinement. This would certainly lead to cruder results. We use *s-ivec* to denote this algorithm in this paper, and Algorithm 3 describes this simplified version.

### 3.4. Segmentation using annotations

In SITW SRC, snippets of hand annotations are given for multi-speaker audio. We exploited this information in the following way.

**Algorithm 3** *s-ivec* diarization

1: **Preparation**
1a. Train UBM, i-vector T matrix using development datasets.
1b. Equally split the features and extract i-vectors for these segments.
2: **Clustering**
2a. PCA, linear transform and dimensionality reduction are performed according to Section 3.3.
2b. Run k-means clustering to obtain the diarization result.

As demonstrated before, all the diarization algorithms start with initial segmentation. Therefore, algorithms treat these annotated parts as known pure segments, and no splits are allowed in the intervals. This should lead to a slightly better initialization. We will discuss the effects later.

### 3.5. Combination of different algorithms

A empirical voting strategy was attempted to combine different diarization algorithms. Frames which get more than $K$ votes from the three algorithms are accepted as the target, i.e., $K = 1$ denotes the union of all results while $K = 3$ means the intersection.

## 4. Experimental work

### 4.1. Experiment setup

The development datasets included NIST SRE 04-08 tel/mic/int excerpts, with Switchboard Phase II Part 1/2/3 and Cellular Part 1/2. There were 34227 male and 45877 female excerpts in total. All UBMs, i-vector matrices, LDA and PLDA models were trained on these data. In the primary speaker recognition system, we used 39-dimension PLP features, 2048-mixture UBM, 400-dimension i-vectors. The i-vector after LDA was 200-dimension.

We used different features for speaker diarization. For BIC-based and aIB diarization, 13-dimension static PLP features with no normalization were used. For s-ivec diarization, both 13-dimension raw and 39-dimension normalized features were evaluated. A 1024-mixture UBM, 100-dimension i-vector T matrix and WCCN transform were trained for 13-dimension features, while a 2048-mixture UBM, 400-dimension i-vector plus 200-dimension LDA were tested for 39-dimension features. We denote them as *s-ivec-1* and *s-ivec-2* respectively. The number of speakers in all multi-speaker audio was assumed to be 2. This assumption was not true, but we have not explored this problem yet. The parameters for the three diarziaions followed common setups in publications [7, 9, 21].

Our system was experimented in SITW speaker recognition database, under *core-core* and *assistclean-core* conditions. The SITW corpus was distributed by SRI in 2016 and was collected from open-source media channels. Hundreds of well-know figures are involved as person of interest. Noise, reverb, various speaking conditions, etc., are introduced, making it close to the real-world situation. Please refer to [15] for more details about SITW database. We report the main metric $C_{det}^{min}$ for both development and evaluation sets in SITW. No calibration problem was considered in this paper.

Table 1: The performance of our system in *assistclean-core* condition. The results are given for $C_{det}^{min}$. The results with or without annotation-assisted segmentation are indicated as *w/o ant.* and *w/ ant.*, respectively.

| Methods | Dev | | Eval | |
|---|---|---|---|---|
| | w/o ant. | w/ ant. | w/o ant. | w/ ant. |
| no diar. | - | 0.6901 | - | 0.6059 |
| ant. only | - | 0.7257 | - | 0.7248 |
| BIC-based | 0.5875 | 0.5878 | 0.5442 | 0.5458 |
| aIB | 0.6031 | 0.5968 | 0.5589 | 0.5451 |
| s-ivec-1 | 0.6014 | 0.6061 | 0.5527 | 0.5427 |
| s-ivec-2 | **0.5787** | **0.5683** | **0.5406** | **0.5288** |

### 4.2. *core-core* condition

The *core-core* condition in SITW SRC is similar to NIST SRE but more challenging. We first validated our system in this condition using the framework introduced in Section 2. The $C_{det}^{min}$ in development (Dev) and evaluation (Eval) datasets are 0.7262 and 0.7457, respectively. These results were achieved without any special technologies or system fusion. Considering the evaluation results of other participants [22], we thought they were quite acceptable.

### 4.3. *assistclean-core* condition

In the *assistclean-core* condition, enrollment audio are relatively clean and one or more speakers are contained in one recording. Diarization and selection modules in Figure 1 were included to find the excerpts of target speakers. Except for the three diarization algorithms, we also tested two trivial methods for comparison. The first one used the whole utterance to train models, while the second one only utilized the annotated snippets. They are abbreviated as *no diar.* and *ant. only* by convenience. The results are shown in Table 1.

From Table 1, we find that *s-ivec-2* diarization consistently achieves the best performance among all algorithms. It is unexpected that *s-ivec-1* performs worse than *s-ivec-2* since it is well known in speaker diarization that raw static features are better and WCCN is more effective than LDA. The inferiority of *s-ivec-1* may due to the smaller UBM and lower dimension i-vectors, but we need more experiments to prove it. We will only discuss *s-ivec-2* below.

If the manual annotations are available, the performance of aIB and s-ivec diarization improves. It is not the case for BIC-based method. We explain the phenomenon that the former two both use equal segmentation, so this auxiliary information may make their initialization more precise.

Table 1 shows the results that the diarization module is very important in the multi-speaker condition. It is too short to achieve a good performance using only the hand annotated snippets, while no-diarization is not a good option neither. The system using s-ivec diarization with annotations(*s-ivec-2*) significantly outperforms the no-diarization one and achieves 16.1% and 12.7% relative $C_{det}^{min}$ improvement on the development and evaluation set, respectively.

Because no speaker transcriptions are available in SITW database, we cannot compute DER for our algorithms. In order to obtain some intuitions between the results of diarization and speaker recognition, we further analyzed the length distribution of split speaker segments. It is presented in Figure 2.

Figure 2 shows that the segments split by s-ivec have the shortest length while the aIB resulted in the longest. The fact seems to indicate some relationship with the results of speaker recognition. Thus, we infer that the s-ivec may achieve the best DER. But what if we only choose some fragments that are more likely spoken by target speakers? Will the speaker recognition system benefits from the shorter segments? This idea still needs more experiments to verify.

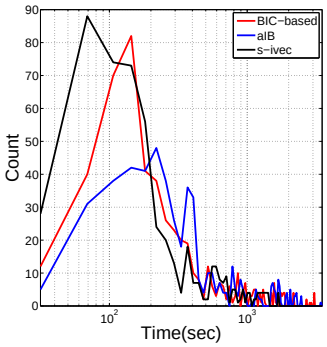


Figure 2: *The length distributions of target speaker segments. The average length for BIC-based, aIB and s-ivec diarization are 446, 567 and 326 seconds, respectively.*

Table 2: The performance of algorithm combination using voting strategy. The results are given for $C_{det}^{min}$..

| | Dev | | Eval | |
|---|---|---|---|---|
| | w/o anno. | w/ anno. | w/o anno. | w/ anno. |
| $K = 1$ | 0.6012 | 0.5887 | 0.5588 | 0.5494 |
| $K = 2$ | **0.5705** | **0.5806** | **0.5393** | **0.5346** |
| $K = 3$ | 0.5954 | 0.5809 | 0.5495 | 0.5433 |

We also evaluated the voting strategy to combine various diarization methods under different values of $K$. Table 2 illustrates the performance. We find that we cannot achieve better results using this strategy. The probable reason is that, when using voting strategy, we introduce bad segments when the results are unified, and miss good segments when intersect them. So new approaches need to be developed to combine the advantages of different diarization algorithms.

## 5. Conclusions

This paper introduces a system framework developed for SITW SRC during the post-evaluation period. Three speaker diarization algorithms are investigated in detail and evaluated in SITW SRC. Experimental results show that, the involvement of speaker diarization greatly improves the performance. Among all the algorithms, the i-vector-based diarization achieves the best results, and the initial manual annotations are helpful to some algorithms. We think the purity of enrollment segments is important in this multi-speaker case.

Future work includes improving the combination method, incorporating a complete version of i-vector diarization and introducing Baum-Welch alignment with deep neural network.

## 6. Acknowledgements

The work is supported by National Natural Science Foundation of China under Grant No. 61370034, No. 61403224 and No. 61273268.

# 7. References


[1] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, May 2014, pp. 1695–1699.

[2] Y. Liu, L. He, and J. Liu, "Improved multitaper pncc feature for robust speaker verification," in *Chinese Spoken Language Processing (ISCSLP), International Symposium on*, 2014, pp. 168–172.

[3] I. Peer, B. Rafaely, and Y. Zigel, "Reverberation matching for speaker recognition," in *Acoustics, Speech and Signal Processing(ICASSP), IEEE International Conference on*, March 2008, pp. 4829–4832.

[4] T. Hasan and J. Hansen, "Acoustic factor analysis for robust speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 842–853, 2013.

[5] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, May 2013, pp. 7649–7653.

[6] Y. Lei and J. Hansen, "The role of age in factor analysis for speaker identification," in *Interspeech*, 2009.

[7] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[8] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[9] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1382–1393, 2009.

[10] D. A. Reynolds, P. Kenny, and F. Castaldo, "A study of new approaches to speaker diarization." in *Interspeech*, 2009, pp. 1047–1050.

[11] O. Kudashev and A. Kozlov, "The diarization system for an unknown number of speakers," in *Speech and Computer: 15th International Conference*, 2013, pp. 340–344.

[12] S. Zhang, C. Zhang, R. Zheng, and B. Xu, "An investigation of summed-channel speaker recognition with multi-session enrollment," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2014, pp. 1640–1644.

[13] H. Sun and B. Ma, "The nist sre summed channel speaker recognition system." in *Interspeech*, 2014, pp. 1111–1114.

[14] C. V. A. Casco, J. V. López, A. O. Giménez, and E. L. Solano, "Speaker verification on summed-channel conditions with confidence measures," *Computación y Sistemas*, vol. 15, no. 1, pp. 27–37, 2011.

[15] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *Submitted to Interspeech*, 2016.

[16] B. Xiang, U. V. Chaudhari, J. Navrátil, G. N. Ramaswamy, and R. A. Gopinath, "Short-time gaussianization for robust speaker verification," in *Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on*, vol. 1, 2002, pp. 681–684.

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

[18] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, 2011, pp. 249–252.

[19] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[20] D. Vijayasenan, F. Valente, and H. Bourlard, "KL realignment for speaker diarization with multiple feature streams," Tech. Rep., 2009.

[21] S. Shum, N. Dehak, E. Chuangsuwanich, D. A. Reynolds, and J. R. Glass, "Exploiting intra-conversation variability for speaker diarization." in *Interspeech*, vol. 11, 2011, pp. 945–948.

[22] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The 2016 speakers in the wild speaker recognition evaluation," in *Submitted to Interspeech*, 2016.