

# Multimodal Clustering with Role Induced Constraints for Speaker Diarization

Nikolaos Flemotomos, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory  
University of Southern California, Los Angeles, CA, USA

flemotom@usc.edu, shri@ee.usc.edu

## Abstract

Speaker clustering is an essential step in conventional speaker diarization systems and is typically addressed as an audio-only speech processing task. The language used by the participants in a conversation, however, carries additional information that can help improve the clustering performance. This is especially true in conversational interactions, such as business meetings, interviews, and lectures, where specific roles assumed by interlocutors (manager, client, teacher, etc.) are often associated with distinguishable linguistic patterns. In this paper we propose to employ a supervised text-based model to extract speaker roles and then use this information to guide an audio-based spectral clustering step by imposing must-link and cannot-link constraints between segments. The proposed method is applied on two different domains, namely on medical interactions and on podcast episodes, and is shown to yield improved results when compared to the audio-only approach.

**Index Terms:** speaker clustering, diarization, speaker roles, spectral clustering, constrained clustering

## 1. Introduction

Speaker diarization is the task of segmenting a multi-party speech signal into speaker-homogeneous regions and tagging them with speaker-specific labels [1, 2]. Even though recently introduced end-to-end neural diarization offers simplicity and achieves remarkable results in some scenarios [3, 4], modular, clustering-based diarization is still widely used and has been an indispensable part of award-winning systems [5, 6]. In the conventional diarization approach, the speech signal is first segmented into chunks assumed to be speaker-homogeneous, in the sense that a single speaker is active therein. Fixed-dimensional speaker representations are then estimated for all the segments and their pairwise similarities are computed. Finally, a clustering algorithm gives the desired labeled speech segments.

Even though it is generally assumed that no information is known a priori about the speakers, in practice we often need to deploy diarization systems in specific applications, and domain-dependent processing can be used to further improve the final performance. To that end, both the acoustic [6] and the linguistic [7] streams of information have been successfully used to either adapt the models or modify the diarization pipeline. The language-based approach, where the transcripts of a recording are taken into consideration during diarization, is especially promising for interactions where speakers play dissimilar roles.

In real-life conversations, interlocutors often assume specific roles which are associated with pre-defined objectives within a group [8]. Those in turn are linked to specific communication patterns which can be manifested through the language used within the conversations. For example, during an interview the interviewer is expected to use more interrogative words than the interviewee and during a flight the air traffic controller is expected to give instructions to the pilot, who responds by reading

back the instructions in a particular manner. It should be noted that several role-playing conversations, such as interviews, clinical interactions, and court hearings, have been included in the evaluation data of recent diarization challenges [9].

This role-specific linguistic variability can be used to improve the performance of core speech processing tasks, including diarization. In our previous work, we used roles to estimate the profiles of the participants in a conversation and reduce the clustering problem into a classification one [7]. In a related work, we ran audio-based speaker clustering and language-based role recognition in parallel and then combined the outputs through a meta-classifier [10]. However, those systems assume that each speaker is linked to a unique role during a conversation. Even though this is often a reasonable assumption (e.g., dialogues between a *clinician* and a *patient*), the systems cannot be easily generalized when a single speaker assumes multiple roles or when multiple speakers play the same role (e.g., trials with a single *judge* and multiple *prosecution witnesses*).

To overcome this limitation, we propose to exploit the linguistically extracted role information to impose constraints during audio-based clustering. Depending on the domain, we can impose must-link and/or cannot-link constraints, without the need for one-to-one correspondence between speakers and roles. Using manually-derived speaker-homogeneous segments with oracle transcriptions, we evaluate the effectiveness of the approach on the clustering performance by running experiments on two domains: i) dyadic clinical interactions, and ii) multi-party interactions from a weekly radio show.

## 2. Background

### 2.1. Spectral clustering for speaker diarization

Clustering is one of the main components in modular speaker diarization. During that step, speech segments are grouped into same-speaker classes, usually following either a hierarchical agglomerative clustering (HAC) [11], or a spectral clustering [12] approach. This grouping is based on the pairwise similarities (e.g., cosine-based) between the  $N$  segments to be clustered, which are stored in an affinity matrix  $\mathbf{W} \in \mathbb{R}^{N \times N}$ .

Having constructed the refined affinity matrix  $\mathbf{W}$  (where refinements are explained later), spectral clustering is a technique that exploits the eigen-decomposition of  $\mathbf{W}$  to project the  $N$  elements onto a suitable lower-dimensional space [13]. To do so, we define the degrees  $d_i \triangleq \sum_j \mathbf{W}_{ij}$  and we construct the normalized Laplacian matrix  $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2}$ , where  $\mathbf{D} = \text{diag}\{d_1, d_2, \dots, d_N\}$ . Assuming we know the number of speakers  $k$ , we find the  $k$  eigenvectors  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$  corresponding to the  $k$  smallest eigenvalues of  $\mathbf{L}$  and form the matrix  $\mathbf{X} = [\mathbf{x}_1 | \mathbf{x}_2 | \dots | \mathbf{x}_k]$ . After normalizing the rows of  $\mathbf{X}$  to unit norm, we cluster them through a  $k$ -means algorithm and assign the  $l$ -th segment to speaker  $s$  iff the  $l$ -th row is assigned to  $s$ .

In order to effectively use spectral clustering in diarization settings, several refinement operations have been proposed [12,

14], the most notable being  $p$ -thresholding. During that step, the  $(100 - p)\%$  largest values in each row of  $\hat{\mathbf{W}}$  are set to 1 and the rest are multiplied by a small constant  $\tau$  (soft thresholding), giving the modified matrix  $\tilde{\mathbf{W}}_p$ . Since this operation may break the symmetry property of the affinity matrix, we re-symmetrize it to get  $\mathbf{W} = \frac{1}{2} (\tilde{\mathbf{W}}_p + \tilde{\mathbf{W}}_p^T)$ .

Instead of fixing a specific value  $p$ , an auto-tuning approach which uses the maximum eigengap of the Laplacian matrix can be followed [14]. The eigengap criterion has its roots in graph theory and is also used to estimate the number of clusters (speakers)  $\hat{k}$ , when this is not known a priori.  $\mathbf{L}$  is a positive semi-definite matrix with  $N$  non-negative real eigenvalues  $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$ . If  $\mathbf{W}$  is viewed as an adjacency matrix of a graph with  $\hat{k}$  perfectly connected components, then  $\hat{k}$  equals the multiplicity of the eigenvalue  $\lambda_1 = 0$ . In practical applications, where we do not expect perfect components,  $\hat{k}$  is estimated by the maximum eigengap as  $\hat{k} = \arg \max_k \frac{\lambda_{k+1}}{\lambda_k}$ .

## 2.2. Constrained spectral clustering

Constrained clustering extends the traditional unsupervised learning paradigm of clustering by integrating supplemental information in the form of constraints [15]. Even though several types of constraints have been explored, the most common ones are the instance-level relations, and in particular the must-link (ML) and cannot-link (CL) constraints. Under that viewpoint, if an ML (CL) constraint is imposed between two segments, then those segments must (must not) be in the same cluster. In speaker diarization, constrained clustering has been applied with constraints imposed either by human input [16], or by acquired knowledge within a particular framework [17–19].

Constraints can be combined with several clustering algorithms, such as k-means [20] or HAC [21]. In this work we use a constrained spectral clustering approach, where constraints are integrated via the exhaustive and efficient constraint propagation (E<sup>2</sup>CP) algorithm [22], which was recently applied in diarization settings [19]. Applying E<sup>2</sup>CP, we can propagate an initial set of pairwise constraints to the entire session. In order to do so, we define a constraint matrix  $\mathbf{Z} \in \mathbb{R}^{N \times N}$ , such that

$$\mathbf{Z}_{ij} = \begin{cases} +1, & \text{if } \exists \text{ ML constraint between } i \text{ and } j \\ -1, & \text{if } \exists \text{ CL constraint between } i \text{ and } j \\ 0, & \text{if } \nexists \text{ any constraint between } i \text{ and } j \end{cases} \quad (1)$$

The elements of the affinity matrix  $\hat{\mathbf{W}}$  are then updated as

$$\hat{\mathbf{W}}_{ij} \leftarrow \begin{cases} 1 - (1 - \mathbf{F}_{ij}^*)(1 - \hat{\mathbf{W}}_{ij}), & \text{if } \mathbf{F}_{ij}^* \geq 0 \\ (1 + \mathbf{F}_{ij}^*)\hat{\mathbf{W}}_{ij}, & \text{if } \mathbf{F}_{ij}^* < 0 \end{cases} \quad (2)$$

where  $\mathbf{F}^*$  contains the constraints propagated to the entire session based on the initial set of constraints and is estimated as

$$\mathbf{F}^* = (1 - \alpha)^2 (\mathbf{I} - \alpha \bar{\mathbf{L}})^{-1} \mathbf{Z} (\mathbf{I} - \alpha \bar{\mathbf{L}})^{-1}. \quad (3)$$

$\bar{\mathbf{L}}$  equals  $\bar{\mathbf{D}}^{-1/2} \hat{\mathbf{W}} \bar{\mathbf{D}}^{-1/2}$ , where  $\bar{\mathbf{D}}$  is a diagonal matrix defined like  $\mathbf{D}$  in Section 2.1, but using the degrees of  $\hat{\mathbf{W}}$ . The constant  $\alpha \in [0, 1]$  is a tunable hyperparameter: a small value penalizes large changes between the initial pairwise constraints in  $\mathbf{Z}$  and the new constraints created during propagation, while a large value penalizes large changes between the neighboring segments in the graph described by  $\hat{\mathbf{W}}$ . Note that for  $\alpha = 0$  we get  $\mathbf{F}^* = \mathbf{Z}$  which means we only rely on the initial constraints, and for  $\alpha = 1$  we get  $\mathbf{F}^* = \mathbf{0}$ , which means we completely ignore any constraint information. The constraint propagation and integration described here takes place before the refinement and spectral operations described in Section 2.1.

## 3. Proposed Method

In this work we propose to use a 2-step clustering for conversations where speakers assume distinct roles (Figure 1).

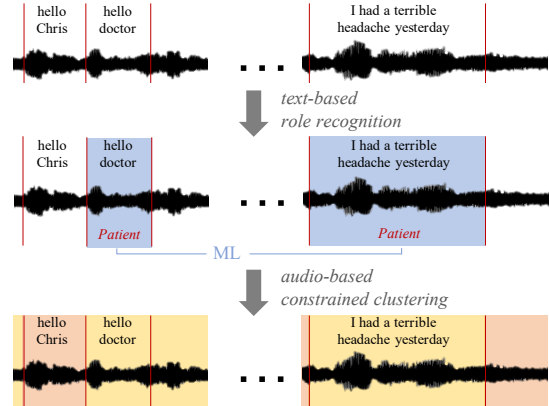


Figure 1: *Two-step speaker clustering for role-playing interactions. Here, an ML constraint is imposed for two segments both associated with the role “patient”. Those segments have to be in the same cluster after the clustering step.*

First, speaker roles are identified from text for each speech segment. To that end, we employ a BERT-based classifier [23], where we add dropout and a softmax inference layer on top of a pre-trained BERT model and we fine-tune it for the task with in-domain data. If, after classification, we have complete role information available (i.e., each segment is associated with a distinct speaker role), we can directly get a purely text-based diarization result. However, there are multiple scenarios where only partial role information is available (e.g., we have sufficient data to only train a binary classifier to identify *news anchor* vs. *guest* in a broadcast news program with multiple potential *guests* within a show). Additionally, we expect that there will be several segments where the linguistic content is not sufficient to robustly infer the associated speaker role.

So, we only use role information to impose suitable constraints for the following step of audio-based clustering, taking into account only segments where roles are identified with sufficient confidence. Even though neural classifiers tend to be overconfident about their decisions and softmax values are usually not a robust proxy of confidence scores, in practice we saw that we can use a softmax threshold as a threshold of confidence, as discussed in Section 5. For those segments where the confidence of their associated role is beyond some specified threshold, we impose ML and CL constraints, according to the domain we are working on. For instance, we can distinguish between the following general scenarios:

1. *different roles are always played by different speakers* (e.g., teacher vs. students): apply a CL constraint between segments associated with different speaker roles,
2. *different speakers always play different roles* (e.g., anchor vs. interviewer vs. guest, where anchor and interviewer might be the same person): apply an ML constraint between segments associated with the same role,
3. *one-to-one correspondence between speakers and roles* (e.g., doctor vs. patient): apply both CL and ML constraints as in cases (1) and (2).

Different types of domain-specific strategies can also be followed. The constraints are then integrated in a spectral clustering algorithm as described in Sections 2.1 and 2.2.

## 4. Datasets

We evaluate the proposed speaker clustering approach on two domains with role-playing interactions. As detailed below, we use a medical dataset from the psychotherapy field and another dataset from the entertainment industry with podcast episodes.

**Psychotherapy sessions.** We use a collection of psychotherapy sessions recorded at a university counseling center (UCC), and specifically the sessions in the sets denoted as  $UCC_{train}$ ,  $UCC_{dev}$  and  $UCC_{test_1}$  in [24]. Each session is a dyadic conversation between a *therapist* and a *patient*, thus falls under case (3) according to the categorization of Section 3. The dataset comprises 97 participants (23 therapists and 74 patients), with no speaker overlap between the train/dev/eval sets. The sessions have been professionally transcribed and the transcribed segments have been forced-aligned with the audio. More details on the dataset are provided in Table 1.

Table 1: Size of the UCC dataset.

	train	dev	eval
#sessions	50	26	20
#segments - therapist	8,766	3,959	4,146
#segments - patient	9,052	4,246	4,245
segment duration (mean)	7.8sec	8.7sec	6.4sec
#words per segment (mean)	21.4	22.3	18.8

**Podcast episodes.** *This American Life*<sup>1</sup> (TAL) is a weekly podcast and radio show, with each episode structured as a storytelling act with multiple characters. The authors in [25] have curated a dataset of 663 TAL episodes aired between 1995 and 2020. We use the audio-aligned utterances provided, with the recommended train/dev/eval split, and with the archived audio standardized as in [25]. Each episode features on average 17.7 speakers (std=8.7) with variable speaking times. The dataset, described in Table 2, has been annotated with speaker IDs and with three speaker roles, those of *host*, *interviewer*, and *subject*. However, the provided role information was not helpful for our purposes, since multiple speakers may play the same role within an episode and, at the same time, a single speaker may play multiple roles (with some episodes having the same speaker occasionally playing all 3 roles). So, we chose to annotate as *host* utterances only the ones spoken by Ira Glass and assign all the other utterances to a *non-host* speaker role. Ira Glass is the host and executive producer of the show and speaks for 18.6% of the time during the entire dataset<sup>2</sup>. Since different roles always denote different speakers (but the inverse does not hold since there are multiple *non-host* speakers), this dataset falls under case (1) according to the categorization given in Section 3.

Table 2: Size of the TAL dataset.

	train	dev	eval
#episodes	593	34	36
#segments - host	26,523	1,765	1,317
#segments - non-host	119,295	6,869	8,039
segment duration (mean)	14.1sec	13.7sec	13.4sec
#words per segment (mean)	37.7	36.6	36.6

<sup>1</sup>[www.thisamericanlife.org](http://www.thisamericanlife.org)

<sup>2</sup>For reference, the second single most-talking speaker of the dataset is Nancy Updike, speaking for 1.6% of the time.

## 5. Experiments and Results

### 5.1. Experimental setup

We run experiments using manually derived speaker segments and transcriptions to avoid propagating potential errors from automated segmentation and speech recognition modules. We standardize the text by stripping punctuation, removing non-verbal vocalizations and converting all letters to lower case.

We build the binary role classifiers (*therapist* vs. *patient* and *host* vs. *non-host*) using TensorFlow [26] with the pre-trained uncased English BERT-base model provided in TensorFlow model garden [27]. Since very short segments are not expected to have sufficient role-related information, during fine-tuning we only take into account segments containing at least 5 words (65.58% of the available training segments for UCC and 88.65% of the available training segments for TAL). We fine-tune the models for 2 epochs, using the Adam optimizer with decoupled weight decay [28], with initial learning rate equal to  $2 \cdot 10^{-5}$  and with a warm-up stage lasting for the first 10% of the training time. The mini-batch size is set to 16 segments and the maximum allowed segment length is set to 128 tokens<sup>3</sup>, which means that 2.91% of the initial training UCC segments and 2.06% of the initial training TAL segments are cropped.

The speaker representation of the segments is based on the widely used x-vectors [29]. To that end, the pre-trained VoxCeleb x-vector extractor from Kaldi [30] is used<sup>4</sup> and a single normalized x-vector is extracted per segment. The segments are clustered following the described constrained spectral clustering approach<sup>5</sup> with cosine-based affinities and with ML and/or CL constraints imposed according to the predicted associated roles. For the UCC dataset, which features dyadic interactions, we group all the segments into two clusters, while for TAL we estimate the number of speakers using the eigengap criterion searching in the range 2–50. The value of  $p$  for  $p$ -thresholding is found through auto-tuning [14], searching in the range 40–95, and we use soft thresholding with  $\tau = 0.01$  [12].

All the results are reported on the eval subsets of the data. Diarization is evaluated with respect to the diarization error rate (DER), estimated with the pyannote.metrics library [31] without allowing any tolerance collar around segment boundaries. DER incorporates three sources of error; false alarms, missed speech, and speaker confusion. However, segmentation is always the oracle one provided by human annotators and, since there is almost no speaker overlap in our datasets, by DER we essentially estimate speaker confusion (false alarm is always 0 and missed speech is 0.02% for UCC and 0.13% for TAL).

### 5.2. Results and discussion

If we have perfect role information for all the segments and if there is one-to-one correspondence between roles and speakers (e.g., one *therapist* vs. one *patient* in UCC), we can get a perfect diarization result in terms of speaker confusion. In the framework of constrained spectral clustering, this can be done by filling  $\mathbf{Z}$  in eq. (1) with all the corresponding constraints and setting  $\alpha = 0$  in eq. (3) so that  $\mathbf{F}^* = \mathbf{Z}$ . This is reflected in Figure 2 where we see how DER changes as we provide more oracle constraints to the algorithm. This is similar to the expected behavior of the algorithm when constraints are added in the form of human supervision.

Without having access to the oracle role information, we

<sup>3</sup>according to the default WordPiece-based BERT tokenizer

<sup>4</sup>[kaldi-asr.org/models/m7](http://kaldi-asr.org/models/m7)

<sup>5</sup>[github.com/wq2012/SpectralCluster](https://github.com/wq2012/SpectralCluster)

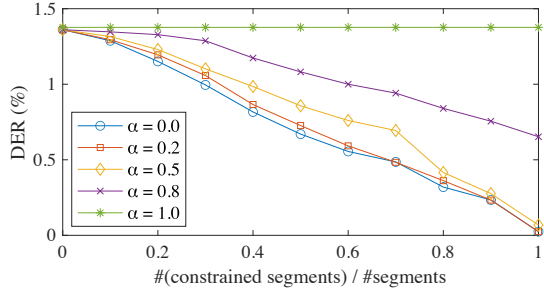


Figure 2: Diarization error rate – lower is better – for the UCC dataset as a function of the (normalized) number of constraints, always providing oracle role information to build the constraints, for different values of  $\alpha$  in eq. (3).

have to rely on a segment-level role classifier. After fine-tuning, the classification accuracy of the BERT-based classifiers, when only segments with at least 5 words are taken into consideration, is 83.22% for UCC (vs. 53.50% for a majority-class baseline) and 90.92% for TAL (vs. 85.41% for a majority-class baseline). Even though the classifiers provide reasonable results, we need to ensure that constraints are imposed only on segments which are confidently linked to some role. For this work, apart from only using segments longer than a specified duration (here, containing at least 5 words) to ensure some minimal linguistic content, we use the softmax values associated with the predicted roles as a proxy of the confidence level.

As shown in Figure 3, the softmax value can indeed act as a reasonable proxy of confidence for our purposes. In particular, if we only consider segments where the corresponding softmax value is above some threshold, accuracy increases monotonically as a function of the threshold. However, the choice of the threshold value is a trade-off decision between accuracy and adequate support so that we have a sufficient number of constraints. For this work, we choose a threshold equal to 0.980 for UCC (accuracy = 94.66%, support = 3,222) and equal to 0.995 for TAL (accuracy = 98.15%, support = 3,674), imposing constraints on around 40% of the segments in both cases.

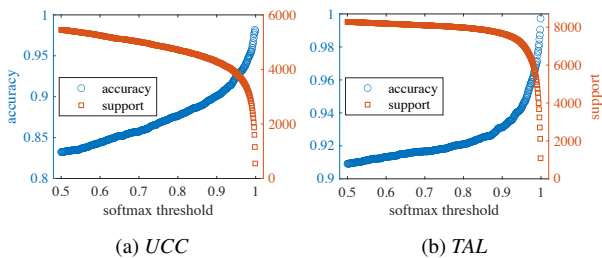


Figure 3: Classification accuracy and support for the BERT-based classifiers when only segments with associated softmax value above some threshold are considered.

After constructing the constraint matrix based on the described role classification only for the segments with sufficient role classification confidence, we perform the constrained spectral clustering algorithm. The results<sup>6</sup> are reported in the second column of Table 3. For comparison, we also provide results for the two extreme cases: i) following a conventional, unconstrained spectral clustering, which ignores any language-based information and ii) following a language-only classification, by using the results of the BERT-based classifier for all the segments, which ignores any audio-based information.

<sup>6</sup>Those results are for  $\alpha = 0.75$  for UCC and  $\alpha = 0.50$  for TAL.

Table 3: Diarization error rate (%) – lower is better – using unconstrained audio-only clustering, constrained clustering with role-induced constraints, and language-only classification.

	unconstrained clustering (audio-only)	constrained clustering (multimodal)	role-based classification (language-only)
UCC	1.38	1.31	10.34
TAL	42.22	23.86	63.01*

\*results contain 2 speakers (due to the binary classification)

In the case of the UCC data, our approach yields a small improvement (5.1% relative) compared to the unconstrained baseline. We additionally found that adding more constraints (selecting a smaller softmax threshold as our confidence criterion) leads to worse performance. Comparing this finding to the results displayed in Figure 2 with oracle constraints, where error approaches 0 given a large number of constraints, we realize that our method is sensitive to the performance of the role classifier. This is because any classification errors can be easily propagated to the clustering step (Figure 1). This error propagation becomes, as expected, more evident in the case we constrain all the segments, relying only on the linguistic stream of information (last column of Table 3).

Looking at the results with the TAL data, we observe a substantial improvement when going from unconstrained to constrained clustering. We can see that in scenarios with a large number of speakers, even partial role-based information can provide useful cues that robustly guide the subsequent clustering. In more detail, we observed that the imposed constraints changed the final Laplacian matrices in a way that the eigen-gap criterion led to the detection of more speakers per episode. The severe performance degradation with the language-only approach is expected, since the results in that case only contain two speakers (since we only have two role classes), even though each TAL episode features multiple participants.

## 6. Conclusion

We proposed to integrate text-based constraints within audio-based clustering to improve the performance of speaker diarization in conversational interactions where speakers assume specific roles. We implemented a BERT-based role classifier solely relying on text data and used its output to construct a constraint matrix for use within constrained spectral clustering. Experimental results in two different domains showed that, after applying a softmax-based confidence criterion, performance can be improved both in cases of one-to-one correspondence between speakers and roles and when only partial role information is available, thus overcoming limitations of previous works.

We performed all our experiments using oracle textual information and oracle speaker segmentation. As part of future work, we intend to study the final performance with respect to the word error rate of a speech recognizer, and to errors due to non-ideal segmentation. Speaker segmentation could be included as a separate module (e.g., [7]), or incorporated within the role recognizer in a named entity recognition (NER)-like approach (e.g., [32]). Future work can also investigate a combination of hard and soft constraints for the task, as well as different types of role-induced constraints. Even though in this work we focused on linguistic characteristics, role-specific behaviors can also be manifest through acoustic, structural, or visual cues, all of which can be potentially used within the framework of role-dependent constrained speaker clustering.

## 7. References

- [1] 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.
- [2] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101317, 2022.
- [3] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Proc. Interspeech*, 2019, pp. 4300–4304.
- [4] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," in *Proc. Interspeech*, 2020, pp. 269–273.
- [5] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny *et al.*, "The STC system for the CHiME-6 challenge," in *Proc. International Workshop on Speech Processing in Everyday Environments (CHIEME)*, 2020, pp. 36–41.
- [6] Y. Wang, M. He, S. Niu, L. Sun, T. Gao, X. Fang, J. Pan, J. Du, and C.-H. Lee, "USTC-NELSLIP system description for DIHARD-III challenge," *arXiv preprint arXiv:2103.10661*, 2021.
- [7] N. Flemotomos, P. Georgiou, and S. Narayanan, "Linguistically aided speaker diarization using speaker role information," in *Proc. The Speaker and Language Recognition Workshop (Odyssey)*, 2020, pp. 117–124.
- [8] A. P. Hare, "Types of roles in small groups: A bit of history and a current perspective," *Small Group Research*, vol. 25, no. 3, pp. 433–448, 1994.
- [9] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The third DIHARD diarization challenge," *arXiv preprint arXiv:2012.01477*, 2021.
- [10] N. Flemotomos, P. Papadopoulos, J. Gibson, and S. Narayanan, "Combined speaker clustering and role recognition in conversational speech," in *Proc. Interspeech*, 2018, pp. 1378–1382.
- [11] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Proc. Interspeech*, 2018, pp. 2808–2812.
- [12] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5239–5243.
- [13] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.
- [14] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2019.
- [15] P. Gañarski, T.-B.-H. Dao, B. Crémilleux, G. Forestier, and T. Lampert, "Constrained clustering: Current and new trends," in *A guided tour of artificial intelligence research*. Springer, 2020, pp. 447–484.
- [16] C. Yu and J. H. Hansen, "Active learning based constrained clustering for speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2188–2198, 2017.
- [17] X. Bost and G. Linares, "Constrained speaker diarization of TV series based on visual patterns," in *IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 390–395.
- [18] K. Kinoshita, M. Delcroix, and N. Tawara, "Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech," in *Proc. Interspeech*, 2021, pp. 3565–3569.
- [19] A. Tripathi, H. Lu, H. Sak, I. L. Moreno, Q. Wang, and W. Xia, "Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) – to appear*, 2022.
- [20] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7198–7202.
- [21] Y. Prokopalo, M. Shamsi, L. Barrault, S. Meignier, and A. Larcher, "Active correction for speaker diarization with human in the loop," in *Proc. Interspeech*, 2021.
- [22] Z. Lu and Y. Peng, "Exhaustive and efficient constraint propagation: A graph-based learning approach and its applications," *International journal of computer vision*, vol. 103, no. 3, pp. 306–325, 2013.
- [23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
- [24] N. Flemotomos, V. R. Martinez, Z. Chen, K. Singla, V. Ardulov, R. Peri, D. D. Caperton, J. Gibson, M. J. Tanana, P. Georgiou *et al.*, "Automated evaluation of psychotherapy skills using speech and language technologies," *Behavior Research Methods*, 2021.
- [25] H. H. Mao, S. Li, J. McAuley, and G. W. Cottrell, "Speech recognition and multi-speaker diarization of long conversations," in *Proc. Interspeech*, 2020, pp. 691–695.
- [26] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "TensorFlow: A system for large-scale machine learning," in *USENIX symposium on operating systems design and implementation (OSDI)*, 2016, pp. 265–283.
- [27] H. Yu, C. Chen, X. Du, Y. Li, A. Rashwan, L. Hou, P. Jin, F. Yang, F. Liu, J. Kim, and J. Li, "TensorFlow Model Garden," <https://github.com/tensorflow/models>, 2020.
- [28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. International Conference on Learning Representations (ICLR)*, 2019.
- [29] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.
- [31] H. Bredin, "pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Proc. Interspeech*, 2017, pp. 3587–3591.
- [32] J. Zuluaga-Gomez, S. S. Sarfjoo, A. Prasad, I. Nigmatulina, P. Motlicek, O. Ohneiser, and H. Helmke, "BERTTraffic: A robust BERT-based approach for speaker change detection and role identification of air-traffic communications," *arXiv preprint arXiv:2110.05781*, 2021.