Speaker Diarization Using Latent Space Clustering in Generative Adversarial Network

Conference Paper · May 2020

DOI: 10.1109/ICASSP40776.2020.9053952

8 authors, including:

Monisankha Pal
University of Southern California
20 PUBLICATIONS **99** CITATIONS
SEE PROFILE

Manoj Kumar
University of Southern California
22 PUBLICATIONS **46** CITATIONS
SEE PROFILE

So Hyun Sophy Kim
Weill Cornell Medical College
41 PUBLICATIONS **577** CITATIONS
SEE PROFILE

Shrikanth S Narayanan
University of Southern California
1,112 PUBLICATIONS **23,311** CITATIONS
SEE PROFILE

SPEAKER DIARIZATION USING LATENT SPACE CLUSTERING IN GENERATIVE ADVERSARIAL NETWORK

Monisankha Pal[1], *Manoj Kumar*[1], *Raghuveer Peri*[1], *Tae Jin Park*[1] *So Hyun Kim*[2], *Catherine Lord*[3], *Somer Bishop*[4], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, University of Southern California
[2]Center for Autism and the Developing Brain, Weill Cornell Medicine
[3]Semel Institute of Neuroscience and Human Behavior, University of California Los Angeles
[4]Department of Psychiatry, University of California, San Francisco

ABSTRACT

In this work, we propose deep latent space clustering for speaker diarization using generative adversarial network (GAN) back-projection with the help of an encoder network. The proposed diarization system is trained jointly with GAN loss, latent variable recovery loss, and a clustering-specific loss. It uses x-vector speaker embeddings at the input, while the latent variables are sampled from a combination of continuous random variables and discrete one-hot encoded variables using the original speaker labels. We benchmark our proposed system on the AMI meeting corpus, and two child-clinician interaction corpora (ADOS and BOSCC) from the autism diagnosis domain. ADOS and BOSCC contain diagnostic and treatment outcome sessions respectively obtained in clinical settings for verbal children and adolescents with autism. Experimental results show that our proposed system significantly outperform the state-of-the-art x-vector based diarization system on these databases. Further, we perform embedding fusion with x-vectors to achieve a relative diarization error rate (DER) improvement of 31%, 36% and 49% on AMI eval, ADOS and BOSCC corpora respectively, when compared to the x-vector baseline using oracle speech segmentation.

***Index Terms*—** ClusterGAN, deep latent space clustering, speaker diarization, speaker embeddings, x-vector

1. INTRODUCTION

Speaker diarization [1], the task of determining "who spoke when" in a multi-speaker audio stream has a wide range of applications from information retrieval and meeting annotations to face to face and telephonic conversation analysis. Recent speaker diarization systems [2, 3] are based on segmenting the input audio stream into uniform speaker-homogeneous segments, followed by extracting fixed-length *speaker embeddings* from those segments and performing speaker clustering over these embeddings.

Among speaker embeddings, i-vectors [4, 5], produced using generative modeling were the first employed for speaker diarization. Recently, embeddings extracted from discriminatively-trained deep neural networks (DNNs) such as d-vectors [6, 7], and *x-vectors* [2, 3] have shown superior performance over i-vectors. These embeddings are partitioned into speaker clusters using clustering algorithms, such as Gaussian mixture models [4], mean-shift [5], agglomerative hierarchical clustering (AHC) [2], k-means [8], spectral clustering [6, 9] and links [10]. All the aforementioned approaches are unsupervised in determining the number of speakers and speaker labels of a given audio session. Recently, a few supervised clustering approaches like UIS-RNN [7] and affinity propagation [11] have also been proposed for diarization.

While performances of tasks such as speech and speaker recognition have improved significantly due to supervised deep learning approaches, most of the existing diarization systems are yet to take full advantage of similar techniques. DNN-based deep clustering approaches are popular in computer vision [12]. While appealing, they are however not immediately applied for speaker diarization tasks probably due to lack of interpretability and the problem of unknown number of speakers of a given audio session. Recently, deep embedded clustering on d-vectors was introduced for speaker diarization [13]. Incorporating the above advances, clustering with dimension reduction using non-linear neural transformation of embeddings, trained with clustering-specific loss could be beneficial for audio diarization systems.

A latent space image clustering method using generative adversarial network (GAN) along with an encoder network (*ClusterGAN*) was proposed recently in [14]. Here, the encoder network performs inverse mapping, i.e., it *back-projects* the data into the latent space. Two main advantages of GAN-based latent space clustering are the interpretability and interpolation in the latent space [14]. In our work, we adopt and modify this network for speaker clustering within the speaker diarization framework. The two main differences of our proposed work from [14] are: (a) instead of random one-hot encoded variables, we use original speaker labels of the training data. Thus, the GAN generator input is a mixture of continuous random and discrete one-hot encoded speaker label variables; (b) instead of images (spectrograms), x-vector embeddings of short audio segments are used as real data input to the GAN discriminator. The GAN and encoder networks are jointly trained along with a clustering-specific loss.

2. BACKGROUND

Over the recent years, the primary focus of research in image clustering has been to non-linearly transform the input feature space to a latent space (where the separation of data is easier) using DNNs. Current deep clustering methods on image data include autoencoder based approaches [15], generative model based approaches such as variational deep embedding [16] and information maximizing GAN (InfoGAN) [17] among others. All these algorithms comprise of three essential components: deep neural network architecture, network loss, and clustering-specific loss. The network loss refers to the reconstruction loss of an autoencoder, variational loss of a variational autoencoder or the adversarial loss of GANs. It is used to learn feasible latent features and avoid trivial solutions. Clustering-specific loss can be cluster assignment losses such as

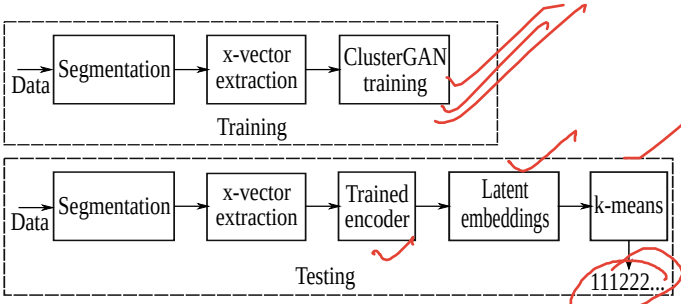

Fig. 1: Schematic diagram of the proposed speaker diarization system.

k-means loss [18], cluster assignment hardening loss [15], spectral clustering loss [19], agglomerative clustering loss [20] or cluster regularization losses such as locality preserving loss, group sparsity loss, cluster classification loss [12]. These losses are used to learn suitable cluster-friendly representations from the data. In this work, we exploit both network loss and clustering loss in the clustering module for speaker diarization.

3. PROPOSED SPEAKER DIARIZATION SYSTEM

3.1. Overview

The overall methodology of the proposed speaker diarization system is shown in Fig. 1. The proposed system begins with the popular time-delay neural network (TDNN) speaker embedding [2], i.e., x-vector extraction and followed by latent space clustering. We discuss each module in the diarization pipeline below.

3.2. Segmentation

Our approach starts with a temporal segmentation of 1.5 sec with 1 sec overlap. The speech segments are embedded into a fixed-dimensional x-vector of dimension 512. This TDNN-based speaker embeddings achieved state-of-the-art performance in speaker verification/diarization [2]. The x-vectors are then fed as inputs to the ClusterGAN network.

3.3. ClusterGAN training

The motivation behind using ClusterGAN on x-vectors is to non-linearly transform it into a lower-dimensional embedding space which is more separable. Although the idea of using a mixture of continuous and discrete latent variables as the input to GAN generator was inspired from InfoGAN [17], ClusterGAN is better suited for clustering than InfoGAN [14]. ClusterGAN comprises three components: the generator (G), the discriminator (D) and the encoder (E), as shown in Fig. 2.

3.3.1. Adversarial training

GANs are a recent class of deep generative models inspired by game theory metaphor, where both G and D networks engage in a two-player minimax game [21]. The generator is considered to be a mapping from the latent space to the data space $G : z \rightarrow \hat{x}$. It takes noisy data z sampled from p_z and generates samples to fool the discriminator. The discriminator is considered to be a mapping from the data space to a real value $D : x \rightarrow \mathbb{R}$. It takes real data x sampled from p_x^r and tries to discriminate between the real and generated fake samples. We employ the improved Wasserstein GAN (IWGAN) [22] for our GAN network. The objective function of this adversarial game is:

$$\min_G \max_D V_{\text{IWGAN}}(D,G) = \mathbb{E}_{x\sim p_x^r}[D(x)] - \mathbb{E}_{z\sim p_z}[D(G(z))] + \lambda \cdot \text{GP} \quad (1)$$

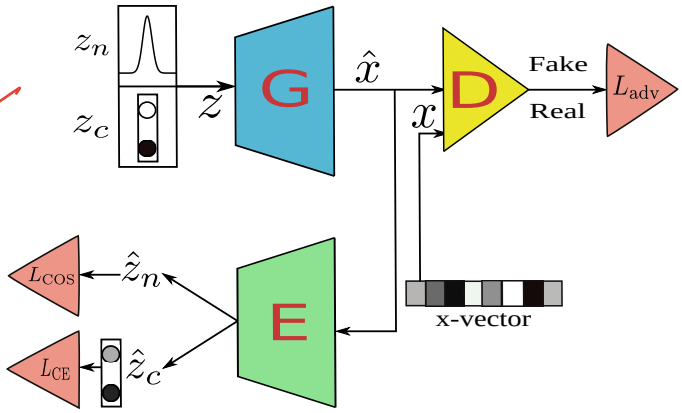

Fig. 2: ClusterGAN architecture. Here, L_{adv}, L_{COS} and L_{CE} represent adversarial, cosine distance and cross-entropy loss functions.

where, λ is the gradient penalty coefficient and GP is the gradient penalty term [22].

3.3.2. Sampling from discrete-continuous mixtures

In order to perform clustering in the latent space, we have to back-project the data into the latent space. The latent space distribution in traditional GANs is typically chosen to be Gaussian or uniform distributions. Although such distributions contain useful information about input data distributions, they usually lead to bad clusters [23]. To mitigate this problem, boosting the latent space using categorical variables to create non-smooth geometry is essential. However, continuity in latent space is also required for good interpolation and GANs have good interpolation ability. Therefore, we employ a mixture of continuous (z_n) and discrete (z_c) variables to the generator by concatenating z_n with z_c. In this work, z_n is randomly sampled from a normal distribution $\mathcal{N}(0, \sigma^2 I_{d_n})$. We chose $\sigma = 0.1$ in all our experiments. We use the original speaker labels for the speech segments from training data as the one-hot encoded variable z_c. The concatenation of z_n with z_c enables clustering in the latent space.

3.3.3. Inverse mapping network

Mapping from the data space to latent space is a non-trivial problem, since it requires the inversion of the generator which is a non-linear model. Existing works [23, 24] tackle this problem by solving an optimization problem in z to get back the latent vectors using $z^* = \text{argmin}_z \mathcal{L}(G(z), x) + \lambda \|z\|_p$, where $\mathcal{L}$ is L_1 norm, λ is a regularization constant and $\|\cdot\|_p$ denotes the norm. However, these approaches are not suitable for clustering since the optimization problem is non-convex [14, 24]. To address this issue, an E network alongside the GAN network for back-projection is introduced. We fix z_c and randomly sample z_n from a normal distribution with multiple restarts at each iteration step. Furthermore, to ensure precise recovery of the latent vector z_n, we compute the numerical difference between the encoder output latent vector $\hat{z}_n$ and z_n. For that, we empirically found that instead of mean square error, cosine distance is more suitable. The objective function for this task can be written as:

$$\min \text{COS}(G,E) = \frac{1}{m}\sum_{i=1}^{m}\left[1 - \frac{E(G(z_n^i)) \cdot z_n^i}{\|E(G(z_n^i))\| \|z_n^i\|}\right] \quad (2)$$

where, m is the mini batch size.

Algorithm 1 ClusterGAN algorithm. Default values: $\lambda = 10$, $m = 64$, $n_{\text{critic}} = 5$, $\alpha = 1\text{e}{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.9$

Require: λ: gradient penalty coefficient; α: learning rate; m: batch size; N_{it}: number of iterations; n_{critic}: number of critic iterations for each generator iteration; α, β_1, β_2: Adam hyper-parameters

for $it = 1$ to N_{it} **do**
  for $\tau = 1$ to n_{critic} **do**
    Sample $\{x^{(i)}\}_{i=1}^{m}$, a batch of x-vectors
    Update the discriminator parameters by

$$\theta \leftarrow \text{Adam}[\nabla_\theta\{\frac{1}{m}\sum_{i=1}^{m} a \cdot [D_\theta(x^{(i)}) - D_\theta(G_\phi(z^{(i)}))+ \lambda \cdot \text{GP}]\}, \theta, \alpha, \beta_1, \beta_2]$$

  end for
  Sample $\{z^{(i)}\}_{i=1}^{m}$, a batch of latent vectors
  Update the generator and encoder parameters by

$$\phi, \psi \leftarrow \text{Adam}[\nabla_{\phi,\psi}\{\frac{1}{m}\sum_{i=1}^{m} -a \cdot D_\theta(G_\phi(z^{(i)}))+ b \cdot \text{COS}(G_\phi, E_\psi) + c \cdot \text{CE}(G_\phi, E_\psi)\}, \phi, \psi, \alpha, \beta_1, \beta_2]$$

end for

3.3.4. *Clustering-specific loss*

To learn cluster friendly representations, we incorporate cluster classification loss while training as cross-entropy (CE) loss. The soft-max layer output obtained by E network is used for computing the cross-entropy loss. This loss encourages the latent embeddings to cluster and hence increase the discriminative information. We minimize this cross-entropy loss as:

$$\min \text{CE}(G, E) = \frac{1}{m}\sum_{i=1}^{m}\left[p(z_{c,i}^{k})\log p\left(E(G(z_{c,i}^{k}))\right)\right] \quad (3)$$

where, the first term is the empirical probability that the embedding belongs to the k-th speaker, and the second term is the predicted probability (by the encoder) that the embedding belongs to the k-th speaker.

3.3.5. *Joint training*

We train the GAN and encoder networks jointly. The training objective function takes the following form:

$$\min_{G,E}\max_{D} [a \cdot V_{\text{IWGAN}}(D, G) + b \cdot \text{COS}(G, E) + c \cdot \text{CE}(G, E)] \quad (4)$$

The weights b and c are used to control the importance of preserving continuous and discrete latent variables. Algorithm 1 shows the training steps of ClusterGAN.

3.4. Testing

After offline training, only the trained encoder model is required to produce the proposed latent embeddings for the input x-vectors of a test audio session. The concatenated latent embeddings (z_n and z_c) are then clustered to produce speaker labels of each segment using k-means.

4. EXPERIMENTAL EVALUATION

4.1. Data preparation

We evaluate our proposed algorithm on the AMI meeting corpus and two child-clinician interaction corpora: ADOS [25] and BOSCC [26]. The AMI database consists of 171 meetings recorded at four different sites (Edinburgh, Idiap, TNO, Brno). For our evaluation, we use the official speech recognition partition of AMI dataset[1]. We exclude the TNO meetings from dev and eval set, which is a common practice in diarization studies [9, 27]. The details of the dataset partition are shown in Table 1.

Table 1: Details of the AMI data set used for our experiments.

	#Meetings	#Speakers
Train	136	155
Dev	14	17
Eval	12	12

The ADOS [28] is a diagnostic tool which comprises over 10 play-based, conversational tasks. We chose two conversational tasks: *Emotions* and *Social Difficulties and Annoyance* from 272 sessions for our evaluation. BOSCC [29] is a new treatment outcome measure, also comprised of play-based, conversational segments. For this study, 24 BOSCC sessions are selected. We use child-clinician datasets from autism research domain to test generalization ability to challenging conditions such as child speech.

4.2. Experimental framework

4.2.1. *Baseline systems*

Since our proposed system uses x-vectors as input features, we used the Kaldi-based AHC clustering with PLDA scoring on x-vectors [2] (denoted as x-vector in this paper) as our main baseline system. We also show results on x-vectors with k-means clustering (denoted as k-means in this paper), as our second baseline.

4.2.2. *Model specifications*

In all our systems, x-vectors are extracted using the Voxceleb[2] models available in the Kaldi recipe. Diarization performance of the proposed system is evaluated for two models trained with different amounts of supervised data: P1 and P2. P1 is trained only on the AMI train set, whereas P2 is trained on both AMI train set and 60 beamformed ICSI [30] sessions with a total number of 46 speakers. The generator and discriminator networks in the proposed systems are simple feed forward neural networks with one and two hidden layers respectively each with 512 nodes. The input layer of G consists of $d = d_n + d_c$ nodes (d_n, d_c are the dimensions of z_n and z_c respectively), where $d_n = 30$ for both P1 and P2 models, and $d_c = 155$ for P1 and 201 for P2 model. G's output layer has 512 nodes, which is the x-vector dimension. The input and output layer of D contains 512 nodes and one node, respectively. On the other hand, the E network consists of a single hidden layer with 512 nodes and input layer is linear with 512 nodes. The output layer of E is a linear layer with d nodes from which the first d_n nodes are directly used as $\hat{z}_n$ and the rest are passed through a soft-max layer to produce $\hat{z}_c$. For all the three networks, the activation function in the hidden layers is ReLU. In the proposed system, we use the original speaker labels from the training data to produce z_c for each segment. The networks are optimized using Adam [31] with a mini-batch size of 64 samples and learning rate $1\text{e}{-4}$. We fixed the a, b and c values as 1, 2 and 10 respectively by tuning on AMI dev set. The number of iterations is fixed to 30k based on optimizing DER on the AMI dev set.

4.2.3. *Performance metrics*

The performance of speaker diarization systems is evaluated by using NIST diarization error rate (DER) [32], which is typically calculated with a 0.25 sec collar. Since the primary focus of this paper is on the effectiveness of new speaker embeddings in clustering, likewise in [2, 9, 27], for all the experiments in this paper we use oracle

[1] http://groups.inf.ed.ac.uk/ami/download/
[2] https://kaldi-asr.org/models/m7

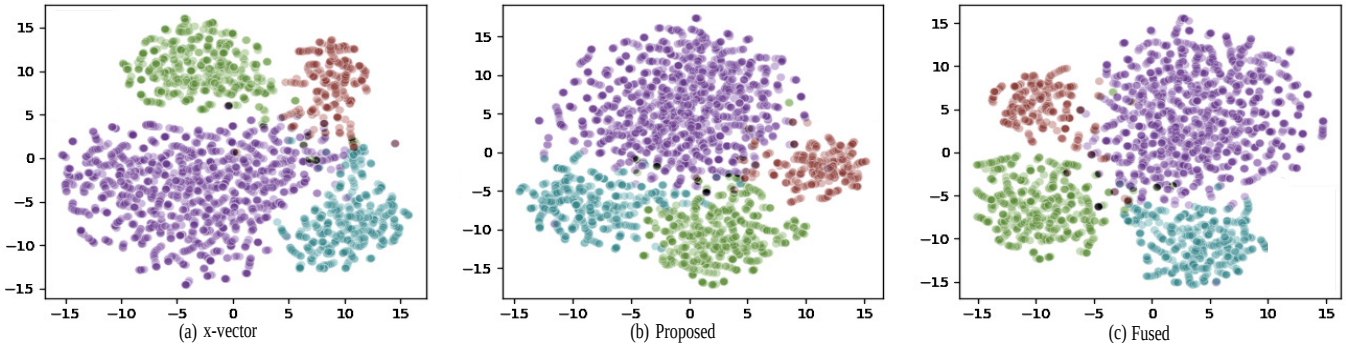

Fig. 3: TSNE visualization of (a) x-vector, (b) proposed and (c) fused embeddings of IS1008a AMI session. This AMI session contains four speakers and each speaker is represented by different colours in the figure.

Table 2: Results on AMI dev and eval set for the baseline and proposed systems.

System	Avg. DER (in %) (oracle SAD, known #speakers)		Avg. DER (in %) (oracle SAD, estimated #speakers)	
	Dev	Eval	Dev	Eval
x-vector	11.65	11.34	11.08	10.37
k-means	11.94	11.45	12.64	12.26
P1	10.17	10.10	10.98	11.26
P2	9.67	11.64	10.33	11.56
x-vector + P1	7.45	**7.82**	8.73	9.11
x-vector + P2	**6.98**	8.85	**7.93**	**8.92**
Sun et. al. [9]	–	–	12.22	12.99

speech activity detection (SAD). Therefore, all DER values reported in this work correspond to speaker confusion errors with no missed or false alarm speech.

4.3. Results and discussions

4.3.1. Results on AMI dev and eval set

Results for diarization performance on AMI dev and eval sets are reported in Table 2. We show results for oracle SAD with both known number of speakers and estimated number of speakers. For the x-vector baseline, we use thresholding on the PLDA scores to perform AHC clustering for unknown number of speakers. The number of speakers for k-means and proposed systems are estimated using Eigen-gap analysis of the affinity matrix constructed from the cosine distance of x-vector embeddings followed by binarization and symmetrization [33]. From Table 2 column 2, we see that for known number of speakers, the P1 system beats x-vector (state-of-the-art) and k-means systems for both AMI dev and eval sets. The performance improves further after incorporating embedding fusion with x-vector embeddings ((x-vector + P1) and (x-vector + P2)). It is observed that both the fused systems significantly outperform all the other systems. The best achieved DER using our fused systems on AMI dev and eval set are 6.98% and 7.82% respectively. This is attributed to the fact that our proposed embeddings have complementary information with x-vector embeddings.

We report the diarization performance of all the systems for estimated number of speakers in Table 2 column 3. Surprisingly, it is observed that x-vector baseline system with thresholding on the PLDA scores for AHC clustering produces a slightly better performance as compared to the oracle number of speaker condition. In contrast, all the other methods' performance degrades for estimated number of speakers. We also compare the proposed diarization system with the work proposed in Sun et al. [9] evaluated on the same data set. The system proposed in [9] is a 2D self-attentive combination of d-vectors with spectral clustering back-end. As seen in Table 2 column 3, our proposed and x-vector fused embeddings with k-means clustering back-end outperforms other baseline methods.

Table 3: Results on ADOS and BOSCC databases for the baseline and proposed systems.

System	Avg. DER (in %) on ADOS	Avg. DER (in %) on BOSCC
x-vector	14.36	21.69
k-means	12.35	14.73
P1	11.27	14.63
P2	11.08	13.35
x-vector + P1	9.38	13.55
x-vector + P2	**9.22**	**11.17**

4.3.2. TSNE visualization

We show TSNE visualizations of x-vector, proposed and fused embeddings of AMI session IS1008a in Fig. 3. It is evident from the figure that the proposed embedding based clusters are slightly more compact as compared to the x-vectors. However, fused embedding based clusters are the most compact within a class and most separated between classes.

4.3.3. Generalization ability

From Table 3, we observe significant performance improvement for the proposed system over the baselines on both ADOS and BOSCC sessions. In addition, the P2 model which is trained on more data achieves better performance than P1 for both individual and fused scenarios. In particular, the improvement is notable compared to the x-vector baseline. We hypothesize that the PLDA model pre-trained on Voxceleb presents a significant domain mismatch in this case. Moreover, both P1 and P2 systems, either used individually or in fusion with x-vectors, are superior to k-means. The best system (x-vector + P2) achieves a relative 36% and 49% improvement over x-vector on those two databases.

5. CONCLUSIONS

We presented a new deep latent space clustering using ClusterGANs to perform speaker diarization. The entire system was trained in a supervised manner along with a clustering-specific loss function. We observed that ClusterGAN-based latent embeddings provide superior performance than x-vector embeddings. Further improvement was achieved after fusing proposed and x-vector embeddings. Experimental results showed a significant DER reduction for the proposed system over state-of-the-art x-vector diarization system on AMI, ADOS and BOSCC corpora. Future work will explore using frame-level representations instead of pre-trained embeddings at the GAN discriminator input, and evaluation on naturalistic settings such as DIHARD.

6. ACKNOWLEDGEMENTS

We gratefully acknowledge the support of USG, NIH and Simons Foundation.

7. REFERENCES

[1] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, 2012.

[2] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree, "Speaker diarization using deep neural network embeddings," in *Proc. ICASSP*, 2017, pp. 4930–4934.

[3] Gregory Sell et al., "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Proc. Interspeech*, 2018, pp. 2808–2812.

[4] Stephen H Shum, Najim Dehak, Réda Dehak, and James R Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2015–2028, 2013.

[5] Mohammed Senoussaoui, Patrick Kenny, Themos Stafylakis, and Pierre Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 1, pp. 217–227, 2014.

[6] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno, "Speaker diarization with LSTM," in *Proc. ICASSP*, 2018, pp. 5239–5243.

[7] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang, "Fully supervised speaker diarization," in *Proc. ICASSP*, 2019, pp. 6301–6305.

[8] Dimitrios Dimitriadis and Petr Fousek, "Developing on-line speaker diarization system," in *Proc. Interspeech*, 2017, pp. 2739–2743.

[9] Guangzhi Sun, Chao Zhang, and Philip C Woodland, "Speaker diarisation using 2D self-attentive combination of embeddings," in *Proc. ICASSP*, 2019, pp. 5801–5805.

[10] Philip Andrew Mansfield, et al., "Links: A high-dimensional online clustering method," *arXiv preprint arXiv:1801.10123*, 2018.

[11] Ruiqing Yin, Hervé Bredin, and Claude Barras, "Neural speech turn segmentation and affinity propagation for speaker diarization," in *Proc. Interspeech*, 2018, pp. 1393–1397.

[12] Elie Aljalbout, Vladimir Golkov, Yawar Siddiqui, Maximilian Strobel, and Daniel Cremers, "Clustering with deep learning: Taxonomy and new methods," *arXiv preprint arXiv:1801.07648*, 2018.

[13] Dimitrios Dimitriadis, "Enhancements for audio-only diarization systems," *arXiv preprint arXiv:1909.00082*, 2019.

[14] Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan, "ClusterGAN: Latent space clustering in generative adversarial networks," in *Proc. AAAI*, 2019, vol. 33, pp. 4610–4617.

[15] Junyuan Xie, Ross Girshick, and Ali Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. ICML*, 2016, pp. 478–487.

[16] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," *arXiv preprint arXiv:1611.05148*, 2016.

[17] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. NIPS*, 2016, pp. 2172–2180.

[18] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong, "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in *Proc. ICML*. JMLR. org, 2017, pp. 3861–3870.

[19] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger, "Spectralnet: Spectral clustering using deep neural networks," *arXiv preprint arXiv:1801.01587*, 2018.

[20] Jianwei Yang, Devi Parikh, and Dhruv Batra, "Joint unsupervised learning of deep representations and image clusters," in *Proc. CVPR*, 2016, pp. 5147–5156.

[21] Ian Goodfellow et al., "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[22] Ishaan Gulrajani et al., "Improved training of Wasserstein GANs," in *Proc. NIPS*, 2017, pp. 5767–5777.

[23] Zachary C Lipton and Subarna Tripathi, "Precise recovery of latent vectors from generative adversarial networks," *arXiv preprint arXiv:1702.04782*, 2017.

[24] Antonia Creswell and Anil Anthony Bharath, "Inverting the generator of a generative adversarial network," *IEEE Trans. on neural networks and learning systems*, 2018.

[25] Daniel Bone et al., "Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist," in *Proc. Interspeech*, 2012.

[26] Manoj Kumar et al., "A knowledge driven structural segmentation approach for play-talk classification during autism assessment.," in *Proc. Interspeech*, 2018, pp. 2763–2767.

[27] Sree Harsha Yella and Andreas Stolcke, "A comparison of neural network feature transforms for speaker diarization," in *Proc. Interspeech*, 2015.

[28] Catherine Lord et al., "The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[29] Rebecca Grzadzinski et al., "Measuring changes in social communication behaviors: preliminary development of the brief observation of social communication change (boscc)," *Journal of autism and developmental disorders*, vol. 46, no. 7, pp. 2464–2479, 2016.

[30] Adam Janin et al., "The ICSI meeting corpus," in *Proc. ICASSP*, 2003, vol. 1, pp. I–I.

[31] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] Jonathan G Fiscus et al., "The Rich Transcription 2006 spring meeting recognition evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 309–322.

[33] Tae Jin Park et al., "The Second DIHARD challenge: System Description for USC-SAIL Team," in *Proc. Interspeech*, 2019, pp. 998–1002.