An Investigation of Speaker Clustering Algorithms in Adverse Acoustic Environments

Meng-Zhen Li*† and Xiao-Lei Zhang*†
* Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen, China
† Center for Intelligent Acoustics and Immersive Communications, School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
E-mail: lemonmengzhen@gmail.com, xiaolei.zhang@nwpu.edu.cn

***Abstract*—Speaker clustering is an important problem of speech processing, such as speaker diarization, however, its behavior in adverse acoustic environments is lack of comprehensive study. To address this problem, we focus on investigating its components respectively. A speaker clustering system contains three components—a feature extraction front-end, a dimensionality reduction algorithm, and a clustering back-end. In this paper, we use the standard Gaussian mixture model based universal background model (GMM-UBM) as a front end to extract high-dimensional supervectors, and compare three dimensionality reduction algorithms as well as two clustering algorithms. The three dimensionality reduction algorithms are the principal component analysis (PCA), spectral clustering (SC), and multilayer bootstrap network (MBN). The two clustering algorithms are the k-means and agglomerative hierarchical clustering (AHC). We have conducted an extensive experiment with both in-domain and out-of-domain settings on the noisy versions of the NIST 2006 speaker recognition evaluation (SRE) and NIST 2008 SRE corpora. Experimental results in various noisy environments show that (i) the MBN based systems perform the best in most cases, while the SC based systems outperform the PCA based systems as well as the original supervector based systems; (ii) AHC is more robust than k-means.**

***Index Terms*—Speaker clustering, noise robust speaker diarization.**

I. Introduction

Speaker clustering aims to partition a set of speech segments into several groups where each group of segments belongs to a single speaker. It is an essential part of many acoustic systems, such as speaker diarization. For example, a speaker diarization system [1] contains four components: data preprocessing, speaker segmentation, speaker clustering and speaker resegmentation. Research experience shows that the performance of speaker diarization is mainly determined by speaker clustering, while data preprocessing and speaker segmentation can be quite standard, and speaker resegmentation does not help much in many cases.

Speaker clustering in clean or high signal-to-noise ratio (SNR) environments has been studied sufficiently. Kenny *et al.* applied agglomerative hierarchical clustering (AHC) in the baseline system [2]. Shum *et al.* [3] used principal component analysis (PCA) for dimension reduction and used k-means clustering based on cosine similarity metric for clustering. They also applied spectral clustering (SC) to identity vectors (i-vectors) [4]. Senoussaoui *et al.* [5] applied an iterative mean-shift algorithm to speaker clustering. Zhang [6] proposed to replace traditional PCA by multilayer bootstrap network (MBN) in a standard speaker clustering system. Wang *et al.* [7] proposed to combine SC with a d-vector based feature extraction front-end, where the SC algorithm contains a novel affinity matrix refinement step. The speaker clustering algorithms in clean environments have achieved good performance.

However, the working environments of speaker clustering in a real-world application are mostly noisy, which needs further study. Recently, this problem has received increasing attention. Zhu *et al.* [8] proposed to enhance noisy speech by a deep neural network based enhancement algorithm, and applied consensus clustering to improve the stability of speaker clustering in noisy conditions. Maciejewski *et al.* [9] built a standard AHC based baseline system and studied speaker diarization in reverberation environments.

In this paper, we aim to study how different speaker clustering systems behave with respect to the variation of signal-to-noise ratio (SNR). The comparison dimensionality reduction algorithms include PCA, SC, and multilayer bootstrap network (MBN) [6], all of which take the supervectors produced from a Gaussian mixture model based universal background model (GMM-UBM) as their inputs. The comparison clustering algorithms include k-means and AHC. We investigate the comparison methods in both in-domain and out-of-domain settings. Experimental results on the noisy versions of NIST 2006 SRE and NIST 2008 SRE corpora show that (i) the MBN based systems perform the best in most cases, while the SC based systems outperform the PCA based systems; (ii) AHC is more robust than k-means.

The rest of the paper is organized as follows. In Section II, we introduce a speaker clustering framework. In Section III, we introduce the nonlinear dimensionality reduction methods in comparison. In Section IV, we report the experimental results. In Section V, we conclude the paper.

II. A speaker clustering framework

A speaker clustering framework that will be shared by all comparison methods is shown in Fig. 1. Suppose there is a set of speech segments $\{\mathbf{t}_i\}_{i=1}^N$, where $\mathbf{t}_i$ is the i-th speech segment in the time domain. The framework first extracts MFCC features from each frame of speech segment $\mathbf{t}$, and then trains a speaker-independent feature extraction front-end, named GMM-UBM, from the pool of all frame-level features

Fig. 1: A speaker clustering framework

[10]. After the GMM-UBM training, it extracts a supervector $\mathbf{z}_i$ which is the zero-th order and first-order Baum-Welch statistics from the i-th speech segment. Finally, speaker clustering becomes a standard clustering problem that partitions $\{\mathbf{z}_i\}_{i=1}^N$ to several non-overlapping groups, each of which belongs to a single speaker.

Because the supervector $\mathbf{z}_i$ is usually high-dimensional and may also contains some nonlinearity, a common idea is to first reduce $\mathbf{z}_i$ to a low-dimensional identity feature vector $\mathbf{x}_i$ by some dimensionality reduction method, such as PCA, SC or MBN, and then conducts clustering on $\{\mathbf{x}_i\}_{i=1}^N$ by traditional AHC or a partition-based method, such as k-means. In speaker diarization and clustering, cosine similarity is suitable to measure the similarity of a pair of identity feature vectors.

III. Dimensionality Reduction Methods

A. Multilayer bootstrap networks

MBN [11] is a nonparametric nonlinear dimensionality reduction method that is flexible in modeling complex and noisy data without pre-assumptions of data distributions, hence it is suitable to the problem of speaker clustering in adverse acoustic environments.

1) Network structure: MBN contains multiple hidden layers and an output layer (Fig. 2). Each hidden layer consists of a group of mutually independent k-centroids clusterings; each k-centroids clustering has k output units, each of which indicates one cluster; the output units of all k-centroids clusterings are concatenated as the input of their upper layer. The output layer is PCA.

The network is gradually narrowed from bottom up, which is implemented by setting parameter k as large as possible at the bottom layer and be smaller and smaller along with the increase of the number of layers until a predefined smallest k is reached.

2) Training method: Before training MBN, we need to first normalize $\mathbf{x}_i$ by its ℓ_2-norm as $\bar{\mathbf{x}}_i \leftarrow \mathbf{x}_i / \|\mathbf{x}_i\|_2$. This preprocessing is to guarantee that, when we evaluate the similarity of $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{x}}_j$ by the inner product $S(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) = \bar{\mathbf{x}}_i^T \bar{\mathbf{x}}_j$, the similarity score equals to the cosine similarity score between $\mathbf{x}_i$ and $\mathbf{x}_j$.

MBN is trained layer-by-layer from bottom up. For training each layer given a d-dimensional input data set $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ either from the lower layer or from the original data space, we simply need to focus on training each k-centroids clustering, which consists of the following steps:

- **Random sampling of features.** The first step randomly selects $\hat{d}$ dimensions of $\mathcal{X}$ ($\hat{d} \leq d$) to form a subset of $\mathcal{X}$, denoted as $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n\}$.

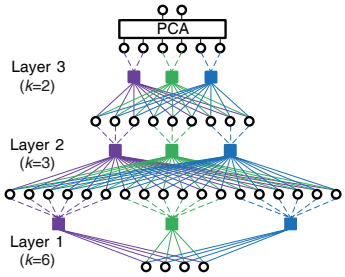

Fig. 2: Network structure [11]. The dimension of the input data for this demo network is 4. Each colored square represents a k-centroids clustering. Each layer contains 3 clusterings. Parameters k at layers 1, 2, and 3 are set to 6, 3, and 2 respectively. The outputs of all clusterings in a layer are concatenated as the input of their upper layer.

- **Random sampling of data.** The second step randomly selects k data points from $\hat{\mathcal{X}}$ as the k centroids of the clustering, denoted as $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$.
- **One-nearest-neighbor learning.** The new representation of an input $\hat{\mathbf{x}}$ produced by the current clustering is an indicator vector $\mathbf{h}$ which indicates the nearest centroid of $\hat{\mathbf{x}}$. For example, if the second centroid is the nearest one to $\hat{\mathbf{x}}$, then $\mathbf{h} = [0, 1, 0, \ldots, 0]^T$. The similarity metric between the centroids and $\hat{\mathbf{x}}$ at the bottom layer is set to $\arg\max_{i=1}^k \mathbf{w}_i^T \hat{\mathbf{x}}$ at all hidden layers.

We summarize the hyperparameters of MBN and their default values in Table I, where the default values are recommended by [11].

B. Spectral analysis of Laplacian matrix

A well-known spectral analysis of Laplacian matrix is SC [12], which conducts eigenvalue decomposition on a normalized Laplacian matrix. It is a widely-used nonparametric nonlinear dimensionality reduction method.

SC first constructs a normalized Laplacian matrix $\mathbf{L}$ by

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2} \tag{1}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{S}$ is an affinity matrix whose element $s_{i,j}$ is the cosine similarity score of $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$s_{i,j} = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}, \quad \forall i, j = 1, \ldots, N \tag{2}$$

and $\mathbf{D}^{-1/2}$ is a diagonal matrix whose diagonal element $d_{i,i}$ is the sum of the i-th row of $\mathbf{W}$.

It then finds k eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ of $\mathbf{L}$ that correspond to the top k smallest eigenvalues, which produces a matrix $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$, where k is the ground-truth number of classes.

TABLE I Hyperparameters of MBN.

Parameter	Description	Default value
δ	A parameter that controls the network structure by $k_{l+1} = \delta k_l$, $\forall l = 1, \ldots, L$	0.5
a	Fraction of randomly selected dimensions (i.e., $\hat{d}$) over all dimensions (i.e., d) of input data.	0.5
V	Number of k-centroids clusterings per layer.	> 100
k_1	A parameter that controls the time and storage complexities of the network.	$k_1 = 0.5N$ for small-scale problems

Suppose the i-th row of $\mathbf{U}$ is $\mathbf{y}_i$, then the feature representation of $\mathbf{x}_i$ produced by SC is:

$$\mathbf{h}_i = \frac{\mathbf{y}_i}{\|\mathbf{y}_i\|_2} \tag{3}$$

IV. Experiments

A. Database

We conducted the investigation on the female speakers of the 8*conv* conditions of the NIST 2006 speaker recognition evaluation (SRE) and NIST 2008 SRE corpora. There are 402 female speakers in the 8*conv* condition of NIST 2006 SRE, and 395 female speakers in the 8*conv* condition of NIST 2008 SRE. Each speaker has 8 utterances. After removing the silence regions by VAD, each utterance is roughly 2 minutes, where we took the ASR transcripts of the corpora as the VAD labels. We cut each utterance into 15 seconds' segments.

We selected babble and factory noises from the NOISEX-92 database as the noise sources. For each type of noise, we took the first two-third part of the noise signal as the noise source of NIST 2006 SRE, and the remaining one-third part as the noise source of NIST 2008 SRE. The experiment was conducted at SNR levels of $[5, 10, 15]$ dB respectively. For each SNR level, we added each speech segment with a randomly selected piece of noise.

Finally, our speaker clustering job is to clustering the noisy speech segments into their ground-truth speakers.

B. Experimental Setup

We adopted two test environments—in-domain test and out-of-domain test. The term "in-domain test" means that the speakers to be clustered have appeared in the GMM-UBM training, while the term "out-of-domain test" means that the speakers are not included in the GMM-UBM training. We used all noisy speech segments of the 402 female speakers in the NIST 2006 SRE to train the GMM-UBM front-end. To simulate a real-world environment, such as meeting or home, we conducted speaker clustering on the first 20 females of NIST 2006 SRE and NIST 2008 SRE as the in-domain and out-of-domain tests, respectively.

We set the frame length to 25 milliseconds and frame shift to 10 milliseconds. We extracted 19-dimensional MFCC with 1-dimensional log energy, and further normalized the 20-dimensional features by feature warping with a 3 seconds sliding window[13]. We trained a GMM-UBM with 1024 Gaussian mixtures for the supervector extraction.

We set the hyperparameters of MBN to the default values in Table I, i.e. $V = 200$, $a = 0.5$, $k_{l+1} = 0.5k_l$, and $k_1 = 0.5N$. The output dimension of PCA was set to the ground-truth number of speakers, i.e. 20. We used k-means (KM) and AHC as the clustering algorithms. Because KM suffers from local minima, we ran KM 50 times and picked the result that corresponds to the optimal objective value among the 50 objective values for each single experiment. We used the unweighted average distance as the clustering metric of AHC. The comparison speaker clustering systems are different combinations of the dimensionality reduction methods and clustering algorithms.

TABLE II Comparison results of speaker clustering systems in the in-domain test and babble noise.

NMI				
	5 dB	10 dB	15 dB	clean
AHC	0.72	0.85	0.89	0.91
SC+AHC	0.55	0.90	0.93	0.94
MBN+AHC	0.72	0.91	0.92	0.93
KM	0.58	0.70	0.75	0.79
SC+KM	0.37	0.87	0.86	0.84
MBN+KM	0.67	0.84	0.82	0.85
ACC				
	5 dB	10 dB	15 dB	clean
AHC	0.56	0.72	0.77	0.79
SC+AHC	0.52	0.85	0.86	0.91
MBN+AHC	0.67	0.91	0.91	0.93
KM	0.46	0.52	0.60	0.65
SC+KM	0.37	0.80	0.71	0.63
MBN+KM	0.56	0.68	0.62	0.69

We evaluated the performance of the speaker clustering systems in terms of normalized mutual information (NMI) and clustering accuracy (ACC), where Hungarian algorithm[1] is used to solve the label permutation problem of ACC between the clustering result and the ground-truth labels. NMI and ACC are two standard evaluation metrics of clustering. The higher their scores are, the better the performance is.

C. Main results

Tables II to V list the results of the comparison speaker clustering systems. From the table, we observe the following experimental phenomena. (i) The performance of all comparison systems drops significantly with the decrease of SNR. (ii) The MBN-based clustering systems are more robust than the SC-based systems and AHC in most noisy environments. For example, MBN+AHC achieves an ACC of 28% higher than AHC and 11% higher than SC+AHC in the in-domain test and factory noise at the SNR of 5 dB. (iii) AHC performs

[1]See http://www.cad.zju.edu.cn/home/dengcai/Data/Clustering.html for the implementation of ACC.

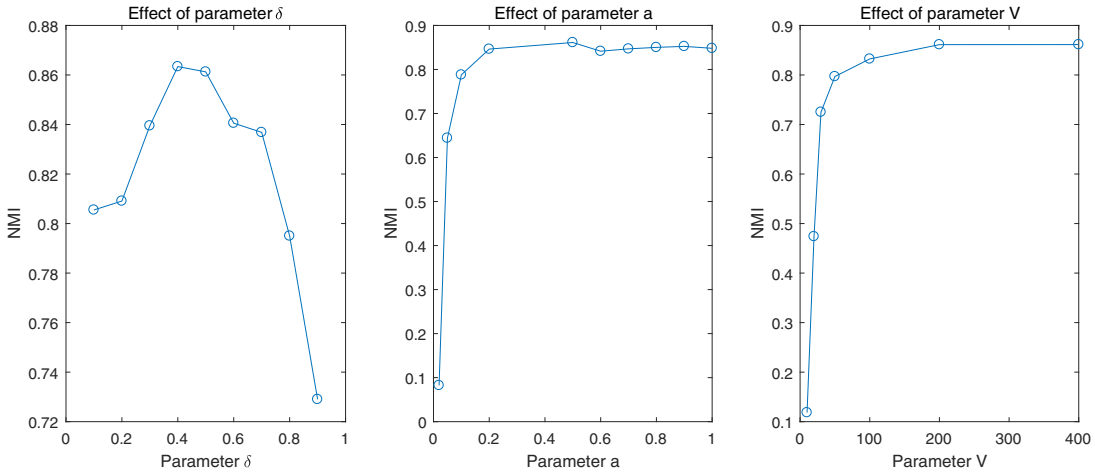

Fig. 4: NMI performance of MBN+AHC with different hyperparameters δ, a, and V in the in-domain test and factory noise at the SNR of 5 dB.

TABLE III Comparison results of speaker clustering systems in the in-domain test and factory noise.

NMI				
	5 dB	10 dB	15 dB	clean
AHC	0.78	0.84	0.86	0.91
SC+AHC	0.83	0.87	0.92	0.94
MBN+AHC	0.86	0.90	0.91	0.93
KM	0.65	0.71	0.77	0.79
SC+KM	0.72	0.82	0.85	0.84
MBN+KM	0.84	0.84	0.82	0.85

ACC				
	5 dB	10 dB	15 dB	clean
AHC	0.60	0.67	0.70	0.79
SC+AHC	0.76	0.76	0.89	0.91
MBN+AHC	0.88	0.91	0.91	0.93
KM	0.55	0.55	0.68	0.65
SC+KM	0.64	0.76	0.70	0.63
MBN+KM	0.82	0.72	0.65	0.69

TABLE IV Comparison results of speaker clustering systems in the out-of-domain test and babble noise.

NMI				
	5 dB	10 dB	15 dB	clean
AHC	0.47	0.53	0.49	0.48
SC+AHC	0.43	0.59	0.58	0.60
MBN+AHC	0.53	0.56	0.58	0.56
KM	0.41	0.44	0.47	0.51
SC+KM	0.38	0.53	0.54	0.56
MBN+KM	0.50	0.52	0.53	0.52

ACC				
	5 dB	10 dB	15 dB	clean
AHC	0.38	0.41	0.34	0.36
SC+AHC	0.37	0.52	0.50	0.51
MBN+AHC	0.45	0.52	0.54	0.52
KM	0.34	0.38	0.39	0.43
SC+KM	0.37	0.45	0.45	0.45
MBN+KM	0.41	0.45	0.42	0.44

TABLE V Comparison results of speaker clustering systems in the out-of-domain test and factory noise.

NMI				
	5 dB	10 dB	15 dB	clean
AHC	0.50	0.52	0.53	0.48
SC+AHC	0.52	0.59	0.58	0.60
MBN+AHC	0.53	0.53	0.56	0.56
KM	0.42	0.42	0.46	0.51
SC+KM	0.44	0.52	0.52	0.56
MBN+KM	0.52	0.54	0.52	0.52

ACC				
	5 dB	10 dB	15 dB	clean
AHC	0.38	0.42	0.39	0.36
SC+AHC	0.44	0.52	0.49	0.51
MBN+AHC	0.49	0.46	0.50	0.52
KM	0.36	0.33	0.41	0.43
SC+KM	0.41	0.47	0.49	0.45
MBN+KM	0.46	0.47	0.44	0.44

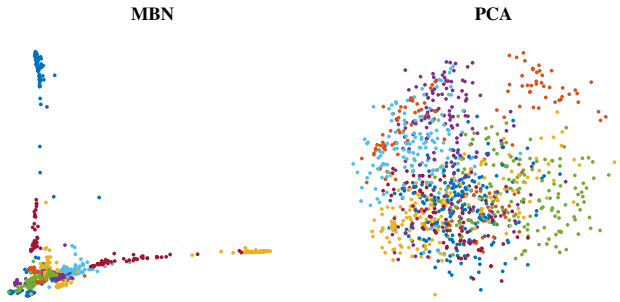

Fig. 3: Visualizations of speaker feature representations produced by MBN and PCA in the in-domain test and factory noise at the SNR of 5 dB. Different colors represent different speakers.

better than KM in most cases. (iv) All systems perform much better in the in-domain test than in the out-of-domain test. An interesting phenomenon is that the speaker clustering methods could not perform well even in the clean environment in the out-of-domain test. (v) The performance of all comparison methods

TABLE VI Performance comparison between MBN-based and PCA-based speaker clustering systems in the in-domain test and factory noise at the SNR of 5 dB.

	NMI	ACC
PCA+AHC	0.81	0.69
MBN+AHC	0.86	0.88
PCA+KM	0.72	0.67
MBN+KM	0.84	0.82

drops significantly in the babble noise at the SNR of 5 dB due to the speech-shaped noise.

We further compared MBN with PCA as a supplemental experiment. Figure 3 shows the visualizations of the 20 speakers in the two-dimensional subspaces produced by MBN and PCA respectively. From the figure, we can see that MBN produces a better visualization than PCA. Table VI lists the comparison results of the MBN-based and PCA-based speaker clustering systems. From the table, we observe that the MBN-based systems outperform the PCA-based systems.

D. Effects of hyperparameters of MBN-based systems on performance

The MBN-based systems, which perform the best in general, have several tunable hyperparameters as shown in Table I. This section studies the hyperparameters that are not default.

We studied the effects of parameters δ, a, V of the MBN+AHC system in the in-domain test and factory noise at the SNR of 5 dB. We searched parameter δ through $[0.1 : 0.1 : 0.9]$, parameter a through $[0.1 : 0.1 : 1]$, and parameter V through $[10, 20, 30, 50, 100, 200, 400]$, where the symbol $[o : p : q]$ means that the search starts at o and ends up at q with an increment of p. The result in Fig. 4 shows that the MBN+AHC with the default parameter setting in Table I reaches nearly the optimal performance.

V. Conclusions

In this paper, we have conducted an empirical comparison between the recent speaker clustering systems in the noisy environments. The motivation behind the work is that the research of previous speaker clustering and diarization is mostly done in the clean environments, however, the real-world working environment of a speaker clustering and diarization method is seldom clean. We have compared several speaker clustering systems, all of which use GMM-UBM as the feature extraction front-end. The systems include AHC, SC+AHC, MBN+AHC, PCA+AHC, KM, SC+KM, MBN+KM, and PCA+KM. We have conducted an in-domain test and an out-of-domain test on the noisy versions of NIST 2006 SRE and NIST 2008 SRE corpora. Experimental results show that (i) the MBN-based speaker clustering systems perform the best in general, while the SC-based systems outperform the systems that use the original supervectors produced from GMM-UBM; (ii) all comparison methods behave much poorer in the out-of-domain test than in the in-domain test; (iii) the performance of all comparison methods drop significantly with the decrease of the SNR levels.

Acknowledgment

This paper was supported in part by the Shenzhen Science and Technology Plan under grant No. JCYJ20170815161820095, in part by the National Natural Science Foundation of China under grant No. 61671381, and in part by the Shaanxi Natural Science Basic Research Program under grant No. 2018JM6035.

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] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[3] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting intra-conversation variability for speaker diarization," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[4] S. Shum, N. Dehak, and J. Glass, "On the use of spectral and iterative methods for speaker diarization," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[5] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 1, pp. 217–227, 2014.

[6] X.-L. Zhang, "Universal background sparse coding and multilayer bootstrap network for speaker clustering." in *INTERSPEECH*, 2016, pp. 1858–1862.

[7] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker Diarization with LSTM," pp. 5239–5243, 2018. [Online]. Available: http://arxiv.org/abs/1710.10468

[8] W. Zhu, W. Guo, and G. Hu, "Feature mapping for speaker diarization in noisy conditions," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5445–5449.

[9] M. Maciejewski, D. Snyder, V. Manohar, N. Dehak, S. Khudanpur, and T. Johns, "CHARACTERIZING PERFORMANCE OF SPEAKER DIARIZATION SYSTEMS ON FAR-FIELD SPEECH USING STANDARD METHODS Center for Language and Speech Processing Human Language Technology Center of Excellence," pp. 5244–5248, 2018.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[11] X.-L. Zhang, "Multilayer bootstrap networks," *Neural Networks*, vol. 103, pp. 29–43, 2018.

[12] 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.

[13] S. O. Sadjadi, M. Slaney, and L. Heck, "Msr identity toolbox v1.0: A matlab toolbox for speaker recognition research," Tech. Rep., September 2013. [Online]. Available: https://www.microsoft.com/en-us/research/publication/msr-identity-toolbox-v1-0-a-matlab-toolbox-for-speaker-recognition-research-2/