Age-Invariant Speaker Embedding for Diarization of Cognitive Assessments

Sean Shensheng Xu[1], *Man-Wai Mak*[1], *Ka Ho Wong*[2], *Helen Meng*[2], *and Timothy C.Y. Kwok*[3,4]

[1]Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University
[2]Dept. of Systems Engineering and Engineering Management, The Chinese University of Hong Kong
[3]Dept. of Medicine and Therapeutics, The Chinese University of Hong Kong
[4]Jockey Club Centre for Osteoporosis Care and Control, The Chinese University of Hong Kong

Abstract

This paper investigates an age-invariant speaker embedding approach to speaker diarization, which is an essential step towards the automatic cognitive assessments from speech. Studies have shown that incorporating speaker traits (e.g., age, gender, etc.) can improve speaker diarization performance. However, we found that age information in the speaker embeddings is detrimental to speaker diarization if there is a severe mismatch between the age distributions in the training data and test data. To minimize the detrimental effect of age mismatch, an adversarial training strategy is introduced to remove age variability from the utterance-level speaker embeddings. Evaluations on an interactive dialog dataset for Montreal cognitive assessments (MoCA) show that the adversarial training strategy can produce age-invariant embeddings and reduce diarization error rate (DER) by 4.33%. The approach also outperforms the conventional method even with less training data.

Index Terms: speaker diarization, Montreal cognitive assessments, age-invariant speaker embedding, deep neural networks

1. Introduction

Cognitive tests are tools for evaluating the cognitive capabilities of humans. There are different types of cognitive tests, such as Montreal cognitive assessments (MoCA) [1], Mini-Mental State Examination (MMSE) [2], and Mini-Cog [3]. Among them, MoCA is a widely used cognitive screening test for detecting mild cognitive impairment (MCI) and Alzheimer's disease in older adults (aged 65 years and above) [1]. One of the most obvious symptoms of MCI and Alzheimer's disease is irregularities in the patient's speech. Therefore, automatic analysis of speech-based MoCA recordings is a useful way to assist specialists in diagnosing early-stage MCI and Alzheimer's disease [4]. To this end, it is essential to develop a speech analysis system that performs automatic speech recognition and determines who spoke when (i.e., speaker diarization) in the cognitive assessments. In this work, we focus on the diarization of MoCA recordings.

Speaker embedding extraction is an essential step for text-independent speaker recognition. Speaker embedding has also been used in speech recognition [5], speech enhancement [6], and speaker diarization [7]. Research [8, 9] on speaker embedding mainly focused on long utterances (over 5 seconds). Nevertheless, speaker embedding on short utterances is also important [10]; this is especially the case for the diarization of MoCA recordings because of the large number of short utterances (less than 1 second) in the interactive dialogs.

Research has found that speaker traits such as age, gender, personality, and voice likability are useful for speaker diarization [11, 12]. In light of this finding, we performed multi-task learning to enrich the age information in the speaker embedding in the early stage of this study. However, because MoCA recordings contain elderly speech, there is an age mismatch between the training data and test data. It turned out that having more age information in the embedding hurts diarization performance instead of helping it. Therefore, in this work, we leverage the idea of domain adversarial neural networks (DANN) [13, 14] to reduce the age variability in the embeddings, which in turn overcomes the age mismatch.

2. Speaker Embedding for Speaker Diarization

In speaker diarization, a conversation involving two or more speakers is divided into a number of overlapping segments. Typically, each segment has a duration of 1.5 seconds, with an overlap of 0.75 seconds. An x-vector extractor [8] is applied to extract a speaker-dependent x-vector for each segment. A PLDA model [15, 16] is then used to compute the pairwise scores of the x-vectors, where each score represents the speaker similarity of two segments. Agglomerative hierarchical clustering [17] or spectral clustering [18] is then applied to the pairwise score matrix to determine the number of speakers and their speech segments' locations within the conversation.

3. Age-Invariant Speaker Embedding with Adversarial Learning

One purpose of this research is to verify the hypothesis that age information in speaker embeddings could be harmful to diarization performance if there is a severe mismatch between the age distributions of training and testing data. If it is the case, adversarial learning can be leveraged to make the embeddings less dependent on age so that the mismatch can be reduced. Age-invariance can be achieved by adding an age classifier and a gradient reversal layer to an utterance-level layer of the x-vector network as shown in Figure 1. The gradient reversal layer keeps the input unchanged during forward propagation but reverses the gradient during backpropagation. This is done by multiplying the gradient with a certain negative constant (the second term in Eq. 1). The reversal of gradient during backpropagation update of the lower part of the x-vector network (below FC6 in Figure 1) assures that the x-vectors are speaker discriminative

This work was in part supported by Research Grands Council of Hong Kong, Theme-based Research Scheme (Ref.: T45-407/19-N).

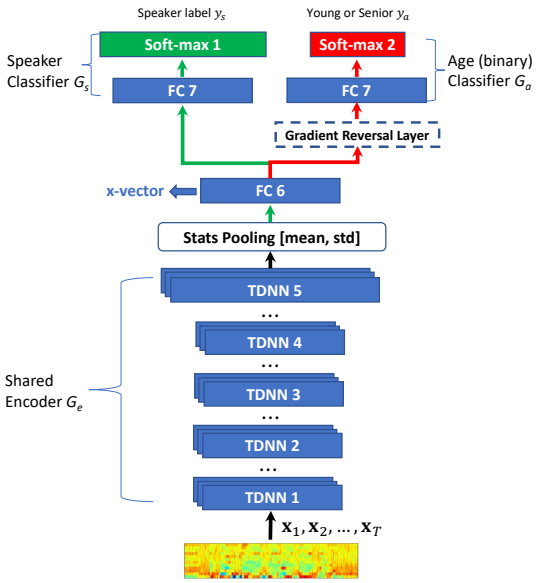


Figure 1: *The x-vector network with adversarial learning. Age-invariant x-vectors can be extracted at layer 6. TDNN and FC represent the time delay neural networks and the fully connected networks, respectively.*

but age confusing, thereby forcing the vectors to be age invariant.

In this work, we applied an x-vector system that uses time-delay deep neural networks (TDNNs) [8] to extract age-invariant speaker embeddings. Figure 1 illustrates the structure of the network, which comprises three parts: shared encoder $G_e(\cdot;\theta_e)$ with parameters θ_e, speaker classifier $G_s(\cdot;\theta_s)$ with parameters θ_s, and age classifier $G_a(\cdot;\theta_a)$ with parameters θ_a. The vector sequence $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ denotes the input of the shared encoder, which are 23-dimensional MFCCs with a frame-length of 25ms and a frame shift of 10 ms. The shared encoder operates at the frame-level, which consists of five time-delay layers. Specifically, assume that t is the current time step; the frames at $\{t-2, \ldots, t+2\}$ are spliced together in the layer TDNN1. In the TDNN2 and TDNN3, we splice together the output of the previous layer at frames $\{t-2, t, t+2\}$ and $\{t-3, t, t+3\}$, respectively. There is no temporal context in the last two layers (TDNN4 and TDNN5). A statistical pooling layer is used to aggregate over the frame-level representation into a feature vector. The frame-level statistics of the input segment are concatenated as the input of a fully connected layer (FC6). Finally, a speaker classifier and an age classifier are added on top of FC6 for speaker classification and age classification, respectively. Rectified linear units (ReLUs) are used as activation functions in the model. The detailed configuration of the network is shown in Table 1.

In Figure 1, the network is trained to optimize the cross-entropy loss $E(\theta_e, \theta_s, \theta_a)$, which is composed of the speaker loss $\mathcal{L}_s$ and the age loss $\mathcal{L}_a$ as follow:

$$E(\theta_e, \theta_s, \theta_a) = \sum_{t=1}^{T} \mathcal{L}_s^t(\theta_e, \theta_s) - \lambda \sum_{t=1}^{T} \mathcal{L}_a^t(\theta_e, \theta_a), \quad (1)$$

where

$$\mathcal{L}_s^t(\theta_e, \theta_s) \equiv \mathcal{L}_s(G_s(G_e(\mathbf{x}_t;\theta_e);\theta_s), y_s), \quad (2)$$

Table 1: *Configuration of the proposed speaker embedding network. T is the segment length (1.5 seconds), N is the number of speakers, and M is the number of outputs in the age classifier. In this work, $M = 1$ because we only have two age groups.*

Layer	Layer Type	Context	In × Out
TDNN1	ConvReLU	$\{t-2, \ldots, t+2\}$	115×512
TDNN2	ConvReLU	$\{t-2, t, t+2\}$	1536×512
TDNN3	ConvReLU	$\{t-3, t, t+3\}$	1536×512
TDNN4	ConvReLU	$\{t\}$	512×512
TDNN5	ConvReLU	$\{t\}$	512×1500
Pooling	Stats	$\{0, \ldots, T-1\}$	1500×3000
FC6	DenseReLU	T	3000×128
FC7	DenseReLU	T	128×128
Softmax1	Dense	T	$128 \times N$
Softmax2	Dense	T	$128 \times M$

$$\mathcal{L}_a^t(\theta_e, \theta_a) \equiv \mathcal{L}_a(G_a(G_e(\mathbf{x}_t;\theta_e);\theta_a), y_a), \quad (3)$$

where y_s and y_a denote the speaker label and age label, respectively. $\mathcal{L}_s$ and $\mathcal{L}_a$ are the standard cross-entropy loss.

To maximize speaker classification performance, instead of using the statistical pooling, a self-attention pooling layer [9] is employed to derive a weighted mean and a weighted standard deviation over each speech segment. We investigated the performance of systems with and without the self-attention mechanism. The results will be shown in Section 5.

4. Experiments

4.1. Training Data

Training data include telephone conversations and interview sessions from the National Institute of Standards and Technology (NIST) Speaker Recognition Evaluations (SREs) and the Switchboard (SWB) datasets. All utterances are in English. The SRE portion consists of SREs from 2004 to 2010. The SWB portion consists of SWB2 Phases 1, 2, and 3 as well as SWB Cellular1 and Cellular2. Table 2 shows the source of data for training the x-vector extractors in different experiments.

The SRE 2008–2010 datasets have over 25,000 telephone conversations and interview sessions spoken by over 1,700 speakers between age 19 and 90. Figure 2 shows the age distributions of the datasets in terms of the number of utterances spoken by these speakers. Two age groups were defined in our experiments: young (under 60) and senior. We can see that the number of utterances of the young group is much larger than that of the senior group. However, in the MoCA data that we used for evaluation (see Section 4.2), each cognitive assessment session comprises a young assessor and a senior. The lack of senior speakers in the training data could make the x-vector extractor biases towards the young speakers, causing unreliable detection of the senior speakers in the evaluation data. In other words, the age mismatch between the training data and the test data will have a detrimental effect on diarization performance, especially for the senior speakers. One solution is to reduce the age information in the speaker embeddings, which motivates us to apply adversarial learning to train the x-vector extractor.

4.2. Evaluation Data

Interactive dialog data collected from the teaching hospital of the Chinese University of Hong Kong (CUHK) were used for evaluation. They were collected for Montreal cognitive assess-

Table 2: *Diarization performance achieved by different systems based on different training data. "with Multi-task Learning" means using multi-task learning (removing the gradient reversal layer in Figure 1) to enrich age information in the speaker embeddings. "with Adversarial Learning" means using adversarial learning (Eqs. 1–3) to make the speaker embeddings age-invariant. "Age Recognition" means the DNN was used to perform age recognition, and the results were obtained by applying the age embeddings. Note that neither multi-task learning nor adversarial learning is applicable to the second column because there is no age information in the training data. DER: diarization error rate; MS: missed speech rate; FA: false alarm rate; SC: speaker confusion rate; att: self-attentive pooling. All performance metrics are in %.*

X-vector extractor Training Data		SRE 2004–2008 + SWB + Augmentation				SRE 2010 + Augmentation				SRE 2008–2010 + Augmentation			
No. of speakers		4979				446				1781			
No. of utterances		62151(clean) 184533(augmented)				15569(clean) 47569(augmented)				25209(clean) 73085(augmented)			
Performance Metrics (%)		DER	MS	FA	SC	DER	MS	FA	SC	DER	MS	FA	SC
without Adversarial Learning	w/o att	7.05	2.5	1.0	3.6	10.59	2.5	1.0	7.1	9.21	2.5	1.0	5.7
	w/ att	**6.39**	2.5	1.0	2.9	10.24	2.5	1.0	6.7	7.85	2.5	1.0	4.4
with Multi-task Learning (Speaker + Age)	w/o att	N/A				19.26	2.4	1.0	15.8	17.03	2.5	1.0	13.5
	w/ att	N/A				19.09	2.4	1.0	15.6	16.92	2.5	1.0	13.4
with Adversarial Learning (Age-invariant Embedding)	w/ att	N/A				**5.91**	2.5	1.0	2.4	**4.92**	2.5	1.0	1.4
Age Recognition (Age Embedding)	w/ att	N/A				14.76	2.4	1.0	11.3	14.04	2.5	1.0	10.5

ments (MoCA), which were used for assessing the patients' cognitive health. Two mobile devices (iPhone 6 and Samsung Galaxy S6) were put in front of the subject on a desk as shown in Figure 3. The recording environment is a quiet office. The collection effort is still ongoing. Each recording consists of a complete Cantonese MoCA test. All assessors and subjects are Cantonese speakers. The age range of subjects is from 72 to 100. Totally, the dataset has 469 conversations, with an average duration of 26 minutes per conversation. This work uses 67 conversations in the dataset.

4.3. Data Augmentation and Network Training

To obtained a robust embedding, the training data were augmented with reverberation, noise, music, and babble. The room impulse responses (RIR) [19] and MUSAN datasets [20] were used for creating room reverberation and additive noise, respectively. The augmentation process roughly doubled the size of the original clean data. Then, the augmented data were combined with the clean data. Table 2 shows the amount of clean and augmented data for training the x-vector extractors. The augmentation was implemented using the Kaldi speech recognition toolkit [21].

We followed the Kaldi's Callhome recipe[1] and used BUT's Tensorflow codes[2] to train the x-vector extractors. For SRE data, we used the Kaldi's energy-based VAD to remove silence regions, whereas, for the MoCA data, we used the ASpIRE SAD.[3]

4.4. PLDA Scoring and Speaker Clustering

We used SRE data (without augmentation) for training PLDA models. We performed the PLDA scoring on all pairs of segments for each recordings. The PLDA scores were then used as input to an agglomerative hierarchical clustering (AHC) algorithm for classifying speech segments by speaker identities. The AHC is an unsupervised clustering and merging method, which has been widely used in speaker diarization systems [22, 23]. Usually, a stopping threshold in the AHC is needed. However, because the number of speakers per recording is known, such stopping threshold is not needed in our case.

4.5. Performance Metrics

Diarization error rate (DER) [24], which is a common performance metric for speaker diarization, was used as performance measure. DER is the sum of missed speech (MS), false alarm (FA) speech, and speaker confusion (SC). Specifically, missed speech error is the percentage of scored time that a speech region is in the reference but not in the hypothesis; false alarm speech error is the percentage of scored time that a speech region is in the hypothesis but not in the reference; speaker confusion error is the percentage of scored time that a segment is assigned to the wrong speaker. A DER of zero indicates perfect diarization and a high DER indicates poorer performance. In our experiments, the measurements were performed using a script named "md-eval.pl" developed by the NIST in the rich transcription (RT) evaluations. To be consistent with other studies [7, 12], a non-scoring collar of 0.25s was used, which refers to the no-score zone around the reference segment boundaries.

5. Results

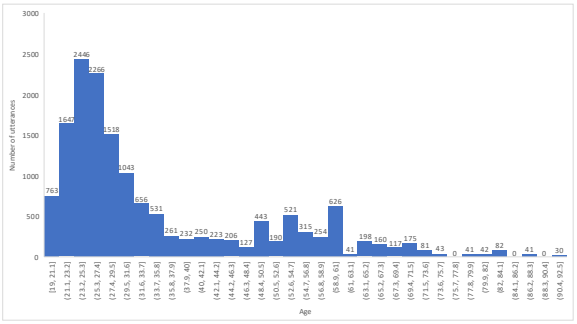
Table 2 shows the DER of the proposed method by applying adversarial learning and conventional embeddings (x-vectors). In the second column, the x-vector extractor was trained using the SRE 2004–2008 and SWB datasets. Results show that a lower DER can be achieved by applying a self-attention mechanism [9]. The training data in the second column are not applicable to age group classification and adversarial learning because no birth information is available in the training datasets.

The datasets (SRE 2008 and SRE 2010) containing the age information of speakers were used to train the x-vector extractors in the third and fourth columns of Table 2. The results show that by combining the data in SRE 2008 and SRE 2010, the

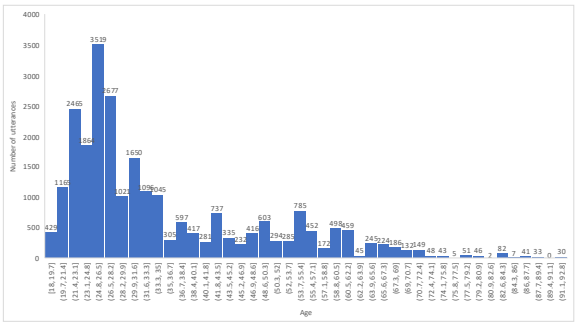
[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/callhome_diarization/v2

[2] https://github.com/hsn-zeinali/x-vector-kaldi-tf

[3] https://kaldi-asr.org/models/m4



(a) SRE 2010



(b) SRE 2008 + SRE 2010

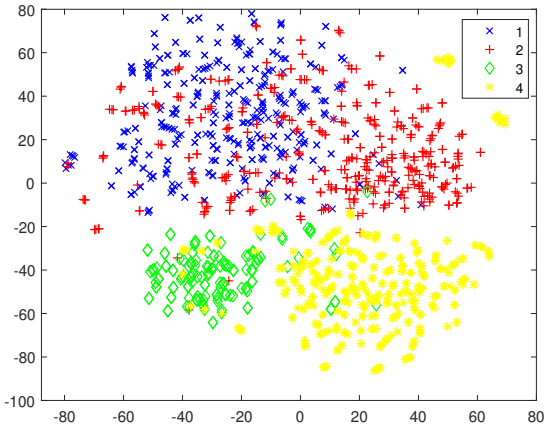
Figure 2: *Age distribution in NIST SRE data.*

DER can be reduced. Using multi-task learning, the DER becomes worse. This suggests that age information in speaker embeddings could be harmful to diarization performance if there is a severe mismatch between the age distributions of training and testing data. Besides, in the last row, the diarization is based on the age information only. The performance of age embeddings is poorer than that of the speaker embeddings. The best results in both third and fourth columns were obtained by using the proposed age-invariant speaker embedding approach. The improvement is significant when compared with the conventional embedding systems. Moreover, for the third and fourth columns, the lowest DERs (i.e., 5.91% and 4.92%) are lower than that of the second column (i.e., 6.39%). This means that the proposed age-invariant diarization works well even though

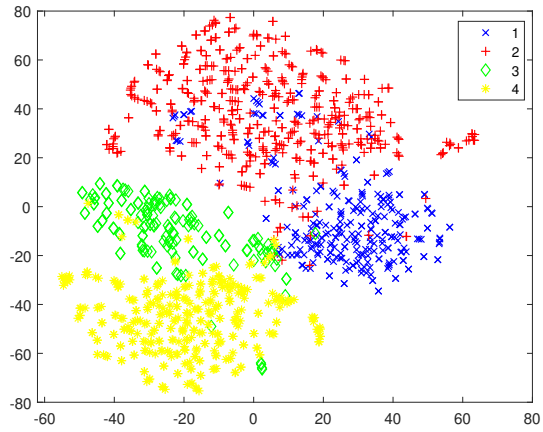
Figure 3: *MoCA data collection.*

the amount of training data has been significantly reduced when only SRE 2008 and SRE 2010 were used.

To show the effect of adversarial learning on the x-vectors, we projected the x-vectors onto two-dimensional t-SNE (t-distributed stochastic neighbor embedding) [25] spaces. The projected speaker embeddings and age-invariant speaker embeddings are shown in Figures 4(a) and (b), respectively. The results show that with adversarial learning, the DNN can generate more discriminative speaker embeddings.



(a) without Adversarial Learning



(b) with Adversarial Learning

Figure 4: *Visualization of utterance-level speaker embeddings by four speakers (1–4). Senior group: blue and green; Young group: red and yellow.*

6. Conclusions

This paper proposes using age-invariant speaker embeddings for the diarization of speech-based MoCA. To make the speaker embeddings invariant to the age mismatch between the speakers in the training and the test data, we leveraged adversarial learning to reduce the age variability. The experimental results show that the detrimental effect of age mismatch is minimized by using the age-invariant embeddings. With the self-attention mechanism, the proposed approach achieves the best diarization performance. Moreover, it outperforms the existing speaker embedding method even with less training data.

7. References

[1] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[2] T. N. Tombaugh and N. J. McIntyre, "The mini-mental state examination: a comprehensive review," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[3] S. Borson, J. Scanlan, and M. B. *et al.*, "The mini-cog: a cognitive "vital signs" measure for dementia screening in multi-lingual elderly," *International Journal of Geriatric Psychiatry*, vol. 15, no. 11, pp. 1021–1027, 2000.

[4] A. Konig, A. Satt, and A. S. *et al.*, "Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people," *Current Alzheimer Research*, vol. 15, no. 2, pp. 120–129, 2018.

[5] W. Li, P. Zhang, and Y. Yan, "TEnet: target speaker extraction network with accumulated speaker embedding for automatic speech recognition," *The Journal of Machine Learning Research*, vol. 55, no. 14, pp. 816–819, 2019.

[6] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. ICASSP 2018*, pp. 5554–5558.

[7] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *Proc. ICASSP 2017*, pp. 4930–4934.

[8] D. Snyder, D. Garcia-Romero, and S. K. D. Povey, "Deep neural network embeddings for text-independent speaker verification," in *Proc. INTERSPEECH 2017*, pp. 999–1002.

[9] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," in *Proc. INTERSPEECH 2018*, pp. 3573–3577.

[10] N. Tawara, A. Ogawa, T. Iwata, M. Delcroix, and T. Ogawa, "Frame-level phoneme-invariant speaker embedding for text-independent speaker recognition on extremely short utterances," in *Proc. ICASSP 2020*, pp. 6799–6803.

[11] Y. Zhang, F. Weninger, B. Liu, M. Schmitt, F. Eyben, and B. Schuller, "A paralinguistic approach to speaker diarisation: Using age, gender, voice likability and personality traits," in *Proceedings of the 2017 ACM on Multimedia Conference*, pp. 387–392.

[12] 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.

[13] Y. Ganin, E. Ustinova, and H. A. *et al.*, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[14] Y. Z. Tu, M. W. Mak, and J. T. Chien, "Variational domain adversarial learning with mutual information maximization for speaker verification," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 2013–2024, 2020.

[15] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV 2007*, pp. 1–8.

[16] M. W. Mak and J. T. Chien, "Machine Learning for Speaker Recognition." Cambridge University Press, 2020.

[17] O. Maimon and L. Rokach, "Data Mining and Knowledge Discovery Handbook." Springer, 2005.

[18] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[19] T. Ko, V. Peddinti, D. Povey, M. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP 2017*, pp. 5220–5224.

[20] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv:1510.08484v1*, 2015.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, and M. H. *et al.*, "The Kaldi speech recognition toolkit," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, 2011.

[22] K. J. Han, S. Kim, and S. S. Narayanan, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1590–1601, 2008.

[23] M. Diez, L. Burget, S. Wang, J. Rohdin, and J. Černocký, "Bayesian HMM based x-vector clustering for speaker diarization," in *Proc. INTERSPEECH 2019*, pp. 346–350.

[24] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*, 1980, pp. 309–322.

[25] L. V. D. Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.