Mathematical Problems of Computer Science 59, 57–68, 2023.
doi: 10.51408/1963-0102

UDC 004.934

Making Speaker Diarization System Noise Tolerant

Davit S. Karamyan[1,2], Grigor A. Kirakosyan[2,3] and Saten A. Harutyunyan[2]

[1]Russian-Armenian University, Yerevan, Armenia
[2]Krisp.ai, Yerevan
[3]Institute of Mathematics of NAS RA, Yerevan, Armenia
e-mail: {dkaramyan, gkirakosyan, sharutyunyan }@krisp.ai

Abstract

The goal of speaker diarization is to identify and separate different speakers in a multi-speaker audio recording. However, noise in the recording can interfere with the accuracy of these systems. In this paper, we explore methods such as multi-condition training, consistency regularization, and teacher-student techniques to improve the resilience of speaker embedding extractors to noise. We test the effectiveness of these methods on speaker verification and speaker diarization tasks and demonstrate that they lead to improved performance in the presence of noise and reverberation. To test the speaker verification and diarization system under noisy and reverberant conditions, we created augmented versions of the VoxCeleb1 cleaned test and Voxconverse dev datasets by adding noise and echo with different SNR values. Our results show that, on average, we can achieve a 19.1% relative improvement in speaker recognition using the teacher-student method and a 17% relative improvement in speaker diarization using consistency regularization compared to a multi-condition trained baseline.

Keywords: Speaker recognition, Speaker diarization, Noise robustness, Teacher-student, Consistency regularization.

Article info: Received 9 January 2023; send to review 30 January 2023, received in revised form 11 April 2023; accepted 17 April 2023.

Acknowledgement: This research was supported by Krisp.ai.

1. Introduction and Related Work

Speaker recognition (SR) is a broad field of study that addresses two major tasks: speaker identification and speaker verification. Speaker identification is the task of identifying a person, whereas speaker verification is the task of determining whether the speaker is who they claim to be. In this study, we focus on far-field, text-independent speaker recognition, where the speaker's identity is determined by the speaking style rather than the content of the speech. Typically, such speaker recognition systems operate on unconstrained speech utterances that are converted into a fixed-length vector known as speaker embedding. Many speech0-processing tasks use speaker embedding such as speaker diarization (SD) [1, 2], automatic speech recognition (ASR) [3], and speech synthesis [4, 5].

In recent years, deep neural networks have actively been employed for speaker embedding extractors since d-vector [6] was proposed. Subsequently, the x-vector [7] was widely used because of the superior performance achieved by employing statistical pooling and time delay neural network (TDNN). Other architectures such as ResNet-based convolutional neural networks and CNNs with cross-convolutional layers [8, 9] were employed for capturing the traits of speech. In addition, to deal with variable-length inputs, Transformer [10], CNN-LSTM [11] and a slew of variants of TDNN [12] were applied for DNN-based speaker embedding extractors. Finally, to reduce the computational complexity and make the models smaller, [13, 14] employed 1D depth-wise separable convolutions for the speaker recognition task.

Metric learning techniques have been successful in speaker recognition tasks. These methods aim to create speaker embeddings with small distances between embeddings of the same speaker and large distances between embeddings of different speakers since unsupervised clustering will be applied to embeddings later in the speaker diarization pipeline. The triplet loss was proposed in [15] which required a careful selection of a triplet because the effectiveness of the performance depended on the contrast between negative and query samples. The prototypical loss was proposed in [16], where many negative samples were used and the Euclidean distance between the centroid of all negative samples and the query embedding was maximized. In the generalized end-to-end loss [17], every utterance in the mini-batch functions as a query as opposed to just one in the prototypical loss. The angular prototypical (AP) loss [18] used only one utterance from each class as the query like the prototypical loss, but with a cosine similarity-based metric.

The primary use case for speaker embeddings is speaker diarization. Speaker diarization is the process of dividing an input audio stream into homogeneous segments according to the speaker's identity. A typical speaker diarization system usually consists of several steps: (1) Speech segmentation, where the input audio is segmented into short sections that are assumed to have a single speaker, and the non-speech sections are filtered out by Voice Activity Detection (VAD), (2) Speaker embedding extractor, where speaker embeddings are extracted from segmented sections, (3) Clustering, where the extracted audio embeddings are grouped [1] into clusters based on the number of speakers present in the audio recording, and optionally, (4) Resegmentation step is performed to further refine clustering results.

In real-world environment, noise causes significant degradations to the performance of speaker diarization systems, and is, hence, a major problem requiring special attention. The goal of noise-tolerant speaker diarization is to achieve improved performance in noisy environments. A recent work [19] tackles this problem using the auto-encoder architecture as a dimensionality reduction module. They extract two low-dimensional codes from speaker embeddings, representing the speaker identity and irrelevant noise information, then remove the noise factors. To our knowledge, there hasn't been a lot of research done in this particular area. ASR systems also suffer deterioration due to audio noise, and this has been the subject of extensive research [20, 21, 22], some of which inspired us.

In this paper, we explore several approaches, borrowed from unsupervised domain adaptation, to make the speaker recognition models noise tolerant. In particular, we apply teacher-student and consistency regularization techniques on speaker recognition and diarization tasks and compare them with multi-condition training when various noise augmentations are used.

We were inspired by the significant results of this work for teacher-student [22], where clean and noisy audios are fed to the teacher and the student, respectively, to enforce similarity between the output distributions. Consistency regularization is a commonly-used technique amongst a variety of tasks in machine learning. This work [20] applies it in a manner similar to that mentioned previously, only here clean and noisy inputs are both fed to the student model. In the paragraphs that follow, we'll discuss in detail how we apply these concepts to obtain noise-robust speaker recognition and diarization.

2. Improving Noise Robustness of Speaker Diarization System

There are several ways to improve the performance of speaker diarization systems in noisy and reverberant environments. For instance, work in [1] proposed the sequence of refinement operations to smooth and denoise data in the similarity space. In this work, we will focus only on the speaker embedding extraction part, and we are going to use unsupervised domain adaptation techniques to make the model noise tolerant.

Given a training dataset consisting of pairs (x_i, y_i) where x_i represents an audio signal and y_i represents the speaker id. Our goal is to learn a parametrized function f_θ, which should be able to compress any given audio into a d-dimensional vector, also known as a speaker embedding. Moreover, if two audio signals are spoken by the same speaker, then the cosine similarity between their corresponding embeddings should be higher. Conversely, if the two audios are spoken by different speakers, the cosine similarity between their embeddings should be lower. The additive angular margin (AAM) loss, as proposed in [23], is a prevalent method for training speaker embedding extractors. The aim of the AAM loss is to minimize the angle between speaker embeddings belonging to the same speaker while simultaneously maximizing the angle between speaker embeddings belonging to different speakers.

2.1 Consistency Regularization

The core idea behind consistency regularization (CR) is to make sure that the network produces similar embeddings for the augmented versions of the same unlabeled utterance [20, 24, 25]. It is enforced by an additional regularization term in the loss function:

$$L_{CR} = \frac{1}{N}\sum_{i=1}^{N} |f_\theta(A(x_i)) - f_\theta(A(x_i))|_2^2,$$

where f_θ is an embedding extractor with parameters θ, N represents the total number of training examples within the dataset. By $A(x)$ we denote a stochastic operation that augments the audio in such a way that the speaker identity remains the same. So the difference is most likely non-zero. The final form of loss is a weighted combination of L_{AAM} and L_{CR} as shown below:

$$L = (1-\alpha)L_{AAM} + \alpha L_{CR},$$

where α is a hyperparameter taking values between 0 and 1.

2.2 Teacher-Student

One critical problem with L_{CR} loss is that it is not stable because of unstable target. To mitigate unstable target problem, the teacher-student model was proposed in [26], where two separate models were used: a Student network with θ parameters and a Teacher with

θ' parameters. On unlabeled examples, the Teacher network provides the learning target for the Student network:

$$L_{TS} = \frac{1}{N}\sum_{i=1}^{N} |f_{\theta}^{Student}(A(x_i)) - f_{\theta'}^{Teacher}(A(x_i))|_2^2.$$

Student is trained as usual. Teacher model is not trained via back-propagation. Instead, its weights are updated at each iteration using the weights from the Student network. Again, the final loss is a weighted combination of $\mathcal{L}_{\mathcal{AAM}}$ and L_{TS} as shown below:

$$L = (1-\alpha)L_{AAM} + \alpha L_{TS}.$$

2.3 Knowledge Distillation

If the teacher model is already trained, it is desirable that its weights remain constant. This training setup is known as ”knowledge distillation”, where the Student model is trained to mimic a pre-trained, larger model [27].

3. Experiments

3.1 Model Architecture

In all experiments, we will use the SpeakerNet [13] architecture as the backbone model. SpeakerNet models are made up of 1D Depth-wise separable convolutional layers. On top of the model, a statistical pooling layer is used to obtain a fixed-length vector. The proposed variation of SpeakerNet (SpeakerNet-M) has fewer parameters (5M) when compared to SOTA and shows very similar performance on VoxCeleb1 [28] trial files when compared to SOTA systems. The model provides embeddings of size 256 for a given audio sample.

In teacher-student experiments, both the teacher and the student have the same architecture.

3.2 Datasets

The VoxCeleb1 [28] and VoxCeleb2 [29] datasets are widely recognized benchmarks in the field of speaker recognition. These datasets have pre-defined development and test sets, which allow for an objective and consistent evaluation of speaker recognition models. We trained our speaker recognition models using only the development part, which consisted of 7205 distinct speakers.

For evaluation of speaker embeddings quality, we use VoxCeleb1 cleaned test trial file. The test trial file contains a list of audio pairs, and the model's performance is evaluated based on its ability to correctly determine whether the two recordings belong to the same speaker or not. To evaluate speaker diarization, we use the VoxConverse [30] development set. The dataset statistics are shown in Table 1.

3.3 Metrics

The equal error rate (EER) metric is used to evaluate the speaker verification. This is the rate used to determine the threshold value for a system when its false acceptance rate and

Table 1: Statistics of datasets used for training SpeakerNet.

Dataset	**# Speakers**	**Duration (h)**	**# Utterances**
VoxCeleb1	1211	340.4	148642
VoxCeleb2	5994	2359.77	1,092,009

false rejection rate are equal. We calculate EER on VoxCeleb1 cleaned test trial file under original, noisy and echo conditions.

For diarization evaluation purposes, we used diarization error rate (DER). This is the sum of three error terms: false alarm (FA), missed detection (MS) and speaker confusion error rate (CER). Similar to the previous works [12, 14], we use collar 0.25 sec and ignore overlap speech regions for confusion error rate calculation. We test the diarization system in original, noisy, and echo scenarios, just like we do for speaker verification.

Both EER and DER are calculated using the cosine similarity back-end.

3.4 Experiment Setup

3.4.1 Input Features

Our audio pre-processing procedure is identical to the one described in the SpeakerNet paper [13]. For each frame window of 20 ms, shifted by 10 ms, 64-dimensional acoustic features were calculated from the speech recordings. Each utterance fed to the encoder has a size $T \times 64$, where T is the number of frames in a given audio sample. We crop speech segments into random chunks from 3 to 8 seconds. With larger chunks, the model converges faster.

3.4.2 Clean Teacher

Our first baseline is a clean teacher trained on VoxCeleb1 and VoxCeleb2 datasets with additive angular margin loss. We set the AAM loss hyperparameters to $s = 30$ and $m = 0.2$, as it was shown in [13, 14], these values give the best results. To avoid overfitting, we added SpecAugment [31] to the training pipeline, which randomly masks blocks of frequency and time channels.

3.4.3 Noisy Teacher

Our second baseline is a noisy teacher trained with the same objective as a clean baseline, and with the additional augmentation steps described below:

- *No Augment*: Leave the utterance unchanged
- *RIR Augment*: Reverberate an input audio using an impulse response from RIRS dataset [32]
- *Noise Augment*: Add noise from MUSAN [33] dataset with signal-to-noise (SNR) values randomly chosen from 0-50DB
- *RIR-Noise Augment*: Apply noise and echo perturbations to the same audio at the same time

- *Speed Augment*: Speed perturbation with 0.95x and 1.05x speeds

RIR, *Noise*, and *RIR-Noise* augmentations all have a probability of 0.25 and are mutually exclusive. *Speed* augmentation is applied independently with a probability of 0.1.

3.4.4 Consistency Regularization

We add an extra mean squared loss between embeddings for the augmented and non-augmented versions of the same utterance to the AAM loss during training.

We set the α hyperparameter in the final loss to 0.1.

3.4.5 Teacher-Student

In order to supervise the student model, we choose our Clean-Teacher baseline as the teacher. We did not update teacher weights during the training and no perturbations were applied to the input of the teacher model. The flow chart of teacher-student training is presented in Fig. 1. During the training procedure, in addition to the AAM loss, the mean squared loss between the student and teacher-produced embeddings is minimized.

We set the α hyperparameter in the final loss to 0.1.

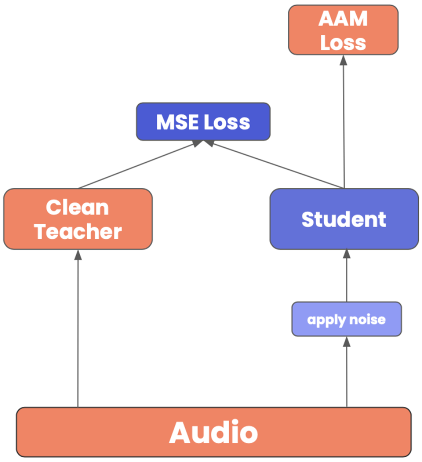

Fig.1. Flow chart of teacher-student learning for improving noise robustness of SR.

3.4.6 Optimization

All models are trained for 200 epochs with an SGD optimizer, with an initial learning rate (LR) of 0.08 using a cosine annealing LR scheduler on 4 A100 GPUs.

3.5 Evaluations

3.5.1 Speaker Verification

All the experiment findings are displayed in Table 2. The results of the original SpeakerNet and the pre-trained checkpoint[1] publicly released by Nvidia are also provided for comparison.

[1] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/speakerverification_speakernet

The pre-trained checkpoint was trained solely with noise augmentation using the above-mentioned datasets. In order to examine the speaker verification system under noisy and reverberant conditions, we created augmented versions of VoxCeleb1 clean test trials by injecting noise and echo with different SNR values.

Table 2: Comparison of different speaker verification models under noise and reverb conditions. The results are reported in equal error rates. The more aggressively noise has been applied, the lower the SNR values were. A noise level of 0 db indicates that the sound and the noise have the same energy.

Model	**Orig**	**0db**	**5db**	**10db**	**Rir**
SpeakerNet [13]	2.14	-	-	-	-
SpeakerNet (NVIDIA)	1.92	9.75	5.43	3.61	16.5
Clean Teacher	1.87	12.9	6.94	4.21	16.5
Noisy Teacher	2.6	9.35	5.84	4.23	12.74
Consistency Reg.	1.76	**8.05**	**4.40**	**3.13**	12.26
Teacher-Student	**1.73**	9.16	4.79	3.26	**9.18**

Table 2 showcases the effectiveness of the methods applied. We can see that training the SpeakerNet model with data augmentation (Noisy Teacher) improves the results in the noisy/reverberant environment with a small deterioration of EER on the original (not perturbed) audios. The Teacher-Student method achieves the lowest EER scores in original and reverberant cases (RIR), whereas the consistency regularization method shows the best results for noisy audios. Using the teacher-student method, we were able to improve the EER by an average of 19.1% compared to the multi-condition trained model. With consistency regularization, we were able to improve the EER by an average of 14.8% compared to the multi-condition trained model.

3.5.2 Speaker Diarization

We employ our trained SpeakerNet models for speaker diarization task to see which model has the smallest performance degradation in noisy conditions. We found that the optimal sliding window size and shift for speech segmentation are 1.5 and 0.5 seconds, respectively. In addition, diarization experiments are based on oracle VAD to evaluate the VAD-independent performance. The affinity matrix A is constructed using the cosine similarity between segment embeddings. We further apply the following sequence of refinement operations to the affinity matrix A:

- *Row-wise Thresholding*: For each row, keep top-12 largest elements and set the rest to 0
- *Symmetrization*: $Y = \frac{1}{2}(A + A^T)$
- *Diffusion*: $Y = AA^T$

We use the spectral clustering method [34] to obtain speaker labels. To get a full picture, we present the diarization results for both known (oracle) and unknown numbers of speakers. In the latter case, we utilize the maximal eigen-gap approach to determine the number of speakers [1].

Table 3: Comparison of speaker diarization systems with various speaker embedding extractors under noise and reverberant conditions. The results are reported in diarization error rate (DER).

Model	**Known #Speakers**						**Unknown #Speakers**					
	0db	**5db**	**10db**	**Rir**	**Orig**	**Avg**	**0db**	**5db**	**10db**	**Rir**	**Orig**	**Avg**
Clean Teacher	12.13	4.48	**1.96**	2.44	**1.26**	4.45	15.44	7.59	2.74	4.48	1.78	6.40
Noisy Teacher	**9.20**	4.49	3.13	3.12	1.57	4.30	**13.09**	7.94	4.18	4.14	1.95	6.26
Consistency Reg.	9.50	3.46	2.0	2.50	1.45	**3.78**	13.40	**4.90**	**2.57**	**3.45**	1.67	**5.20**
Teacher-Student	9.84	**3.41**	2.11	**2.43**	1.36	3.83	13.99	6.17	3.09	3.52	**1.61**	5.67

In order to assess the performance of the speaker diarization system under noisy and reverberant conditions, we modified the Voxconverse dev dataset by adding noise and echo at various signal-to-noise ratios. The results, shown in Table 3, indicate that the teacher-student and consistency regularization methods generally outperform the multi-condition baseline model for both scenarios involving known and unknown numbers of speakers. In particular, when the number of speakers is unknown, we observed approximately 17% and 9.5% relative performance improvements for the consistency regularization and teacher-student methods, respectively, compared to the multi-condition baseline.

However, it is worth noting that in certain specific scenarios, the baseline models may outperform the models with the overall best average performance.

4. Conclusions

In this research, we explore ways to increase the accuracy of speaker recognition and speaker diarization in noisy and reverberant environments, such as multi-condition, teacher-student, and consistency regularization. The key component of the methods used is the additional regularization term between embeddings for augmented and non-augmented versions of the same utterance. Through the use of teacher-student and consistency regularization, we were able to improve the performance of SpeakerNet on speaker recognition and diarization tasks in noisy and reverberant situations.

References

[1] Q. Wang, C. Downey, L. Wan, P. Mansfield and I. Moreno, “Speaker diarization with LSTM”, *2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 5239-5243, 2018.

[2] 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, pp. 356-370, 2012.

[3] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. Saurous, R. Weiss, Y. Jia, and I. Moreno, “Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking”, *ArXiv Preprint ArXiv:1810.04826*, 2018.

[4] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, and Others, “Transfer learning from speaker verification to multi-

speaker text-to-speech synthesis", *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[5] E. Cooper, C. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings", *ICASSP 2020-2020 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*, pp. 6184-6188, 2020.

[6] E. Variani, X. Lei, E. McDermott, I. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification", *2014 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*, pp. 4052-4056, 2014.

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition", *2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*, pp. 5329-5333, 2018.

[8] Y. Yu, L. Fan, and W. Li, "Ensemble additive margin softmax for speaker verification", *ICASSP 2019-2019 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*, pp. 6046-6050, (2019).

[9] Z. Gao, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An improved deep embedding learning method for short duration speaker verification", International Speech Communication Association, 2018.

[10] P. Safari, M. India, and J. Hernando, "Self-attention encoding and pooling for speaker recognition", *ArXiv Preprint ArXiv:2008.01077*, 2020.

[11] J. Jung, H. Heo, I. Yang, H. Shim, and H. Yu, "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result", *IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5349-5353, 2018.

[12] N. Dawalatabad, M. Ravanelli, F. Grondin, J.Thienpondt, B. Desplanques and H. Na, "ECAPA-TDNN embeddings for speaker diarization", *ArXiv Preprint ArXiv:2104.01466*, 2021.

[13] N. Koluguri, J. Li, V. Lavrukhin and B. Ginsburg, "SpeakerNet: 1D depth-wise separable convolutional network for text-independent speaker recognition and verification", *ArXiv Preprint ArXiv:2010.12653*, 2020.

[14] N. Koluguri, T. Park and B. Ginsburg, "TitaNet: Neural Model for speaker representation with 1D Depth-wise separable convolutions and global context", *Proceedings of the IEEE International Conference on Acoustics, Speech And Signal Processing (ICASSP)*, pp. 8102-8106, 2022.

[15] F. Schroff, D. Kalenichenko and J. Philbin, "Facenet: A unified embedding for face recognition and clustering", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815-823, 2015.

[16] J. Snell, K. Swersky and R.Zemel, "Prototypical networks for few-shot learning", *Advances in Neural Information Processing Systems*, vol.30, 2017.

[17] L. Wan, Q. Wang, A. Papir and I. Moreno, "Generalized end-to-end loss for speaker verification", *Proceedings of the IEEE International Conference on Acoustics, Speech And Signal Processing (ICASSP)*, pp. 4879-4883, 2018.

[18] J. Chung, J. Huh, S. Mun, M. Lee, H. Heo, S. Choe, C. Ham, S. Jung, B. Lee and I. Han, “In defence of metric learning for speaker recognition”, *ArXiv Preprint ArXiv:2003.11982*, 2020.

[19] Y. Kim, H. Heo, J. Jung, Y. Kwon, B. Lee and J. Chung, “Disentangled dimensionality reduction for noise-robust speaker diarization”, *ArXiv Preprint ArXiv:2110.03380*, 2021.

[20] Y. Hu, N. Hou, C. Chen E. Chng, “Dual-path style learning for end-to-end noise-robust speech recognition”, *ArXiv Preprint ArXiv:2203.14838*, 2022.

[21] Q. Zhu, J. Zhang, Z. Zhang, M. Wu, X. Fang and L. Dai, “A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition”, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3174-3178, 2022.

[22] L. Moner, M. Wu, A. Raju, S. Parthasarathi, K. Kumatani, S. Sundaram, R. Maas, and B. Hoffmeister, “Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning”, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6475-6479, 2019.

[23] J. Deng, J. Guo, N. Xue and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition”, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690-4699, 2019.

[24] A. Vanyan and H. Khachatrian, “Deep semi-supervised image classification algorithms: a survey”, *J. Univers. Comput. Sci.*, vol. 27, pp. 1390-1407, 2021.

[25] S. Laine and T. Aila, “Temporal ensembling for semi-supervised learning”, *ArXiv Preprint ArXiv:1610.02242*, 2016.

[26] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results”, *Advances in Neural Information Processing Systems*, vol.30, 2017.

[27] G. Hinton, O. Vinyals and J. Dean, “Distilling the knowledge in a neural network. *ArXiv Preprint ArXiv:1503.02531*, 2015.

[28] A. Nagrani, J. Chung and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset”, *ArXiv Preprint ArXiv:1706.08612*, 2017.

[29] J. Chung, A. Nagrani and A. Zisserman, “Voxceleb2: Deep speaker recognition”, *ArXiv Preprint ArXiv:1806.05622*, 2018.

[30] J. Chung, J. Huh, A. Nagrani, T. Afouras and A. Zisserman, “Spot the conversation: speaker diarization in the wild”, *ArXiv Preprint ArXiv:2007.01216*, 2020.

[31] D. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. Cubuk and Q. Le, “Specaugment: A simple data augmentation method for automatic speech recognition”, *ArXiv Preprint ArXiv:1904.08779*, 2019.

[32] T. Ko, V. Peddinti, D. Povey, M. Seltzer and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition”, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220-5224, 2017.

[33] D. Snyder, G. Chen and D. Povey, “Musan: A music, speech, and noise corpus”, *ArXiv Preprint ArXiv:1510.08484*, 2015.

[34] U.Von Luxburg, “A tutorial on spectral clustering”, *Statistics and Computing*, vol. 17, pp. 395-416, 2007.

Աղմկադիմացկունության ապահովումը խոսնակների դիարիզացիայի համակարգում

Դավիթ Ս. Քարամյան[1,2], Գրիգոր Ա. Կիրակոսյան[2,3], Սաթեն Ա. Հարությունյան[2]

[1]Հայ-Ռուսական համալսարան, Երևան, Հայաստան
[2]Krisp.ai, Երևան, Հայաստան
[3]ՀՀ ԳԱԱ մաթեմատիկայի ինստիտուտ,Երևան, Հայաստան
e-mail: {dkaramyan, sharutyunyan, gkirakosyan }@krisp.ai

Ամփոփում

Խոսնակների դիարիզացիայի նպատակը աուդիո ձայնագրության մեջ տարբեր խոսնակների հայտնաբերումն ու առանձնացումն է: Այնուամենայնիվ, ֆոնային աղմուկը կարող է ազդել այս համակարգերի ճշգրտության վրա: Այս հոդվածում ուսումնասիրվել են այնպիսի մեթոդներ, ինչպիսիք են՝ տարբեր աուգմենտացիաներով ուսուցումը, կայունության կարգավորումը (consistency regularization) և ուսուցիչ-աշակերտ մեթոդը՝ խոսնակների ձայնային հատկանիշներ դուրս բերող մոդելի կայունությունը աղմուկի նկատմամբ բարձրացնելու համար: Նշված մեթոդների արդյունավետությունը ստուգվել է խոսնակների նույնականացման և դիարիզացիայի խնդիրներում և ցույց է տրվել, որ դրանք հանգեցնում են կայունության բարելավմանը՝ աղմուկի և արձագանքի առկայության դեպքում: Խոսնակների նույնականացման և դիարիզացիայի համակարգերը աղմուկի և արձագանքի պայմաններում փորձարկելու համար ստեղծվել են VoxCeleb1 և Voxconverse dev տվյալների հավաքածուների ընդլայնված տարբերակները՝ ավելացնելով տարբեր SNR արժեքներով ֆոնային աղմուկ և արձագանք: Ստացված արդյունքները ցույց են տալիս, որ միջին հաշվով կարելի է հասնել խոսնակների նույնականացման ճշգրտության հարաբերական բարելավմանը՝ $19,1\%$-ով՝ օգտագործելով ուսուցիչ-աշակերտ մեթոդը և խոսնակների դիարիզացիայի ճշգրտության հարաբերական բարելավմանը՝ 17%-ով՝ օգտագործելով կայունության կարգավորման մեթոդը՝ համեմատած տարբեր աուգմենտացիաներով վարժեցված մոդելի հետ:

Բանալի բառեր՝ խոսնակների նույնականացում, խոսնակների դիարիզացիա, աղմկա-դիմացկունություն, ուսուցիչ-աշակերտ, կայունության կարգավորում:

Обеспечение шумоустойчивости системы диаризации дикторов

Давид С. Карамян[1,2], Григор А. Киракосян[2,3], Сатен А. Арутюнян[2]

[1]Российско-Армянский университет, Ереван, Армения
[2]Krisp.ai, Ереван, Армения
[3]Институт математики НАН РА, Ереван, Армения
e-mail: dkaramyan, sharutyunyan, gkirakosyan@krisp.ai}

Аннотация

Целью системы диаризации дикторов является идентифицирование и разделениеразных дикторов в аудиозаписи. Однако шум в записи может повлиять на точность этих систем. В этой статье мы исследуем такие методы, как обучение с различными аугментациями, регуляризация согласованности (consistency regularization) и метод "учитель-ученик", чтобы повысить устойчивость экстракторов речевых характеристик к шуму. Мы проверяем эффективность этих методов в задачах распознавания дикторов по голосу и диаризации дикторов и демонстрируем, что они приводят к улучшению устойчивости при наличии шума и реверберации. Чтобы проверить систему распознавания и диаризации дикторов в условиях шума и реверберации, мы создали расширенные версии VoxCeleb1 и наборов данных Voxconverse dev, добавив шум и эхо с разными значениями SNR. Наши результаты показывают, что в среднем мы можем добиться относительного улучшения распознавания дикторов на $19,1\%$ с использованием метода "учитель-ученик" и относительного улучшения диаризации дикторов на 17% с использованием метода регуляризации согласованности по сравнению с базовой моделью, обученной с помощью различных аугментаций.

Ключевые слова:распознавание по голосу, диаризация дикторов, устойчивость к шуму, учитель-ученик, регуляризация согласованности.