RETHINKING SESSION VARIABILITY: LEVERAGING SESSION EMBEDDINGS FOR SESSION ROBUSTNESS IN SPEAKER VERIFICATION

Hee-Soo Heo[1], KiHyun Nam[2], Bong-Jin Lee[1], Youngki Kwon[1], Minjae Lee[1], You Jin Kim[1], Joon Son Chung[2]

[1]NAVER Cloud Corporation, South Korea
[2]Korea Advanced Institute of Science and Technology, South Korea

ABSTRACT

In the field of speaker verification, session or channel variability poses a significant challenge. While many contemporary methods aim to disentangle session information from speaker embeddings, we introduce a novel approach using an additional embedding to represent the session information. This is achieved by training an auxiliary network appended to the speaker embedding extractor which remains fixed in this training process. This results in two similarity scores: one for the speakers information and one for the session information. The latter score acts as a compensator for the former that might be skewed due to session variations. Our extensive experiments demonstrate that session information can be effectively compensated without retraining of the embedding extractor.

Index Terms— Speaker verification, speaker embedding, session information

1. INTRODUCTION

In the evolving domain of speech processing, speaker verification plays a crucial role, having various real-world applications ranging from voice-based security systems to personalised speech assistants. Central to robust speaker verification is the extraction of speaker embeddings, which encapsulate the unique characteristics of an individual's voice [1–3]. However, these embeddings are susceptible to extraneous information, largely influenced from the recording environment. Variabilities in recording devices, ambient noise, room acoustics, and other session-related factors can significantly affect the accuracy of these embeddings, creating misleading similarities even among distinct speakers in similar recording situations [4, 5].

Historically, when the i-vector approach was prevalent in the speaker embedding space, techniques such as linear discriminant analysis (LDA) and within-class covariance normalization (WCCN) were employed as countermeasures to diminish these unexpected session similarities [1, 6]. With the advances of deep learning and its application to this domain, efforts have shifted towards disentangling speaker information from session information directly within the embedding [5, 7, 8]. Various strategies have been studied in this direction – while some leverage the adversarial approach, others design novel loss functions to achieve the same goal [9]. However, a clear problem with these methods is that while trying to separate session-related information from speaker-specific details, important characteristics of the speaker might be lost. In simpler terms, in the process of removing unwanted session information, one might also unintentionally remove features that help identify the speaker.

In light of these challenges, this paper introduces an alternative approach. Instead of disentangling session-related information from the embedding, we present a framework to compensate for it at the score level. Our methodology capitalises on the use of an auxiliary network, seamlessly appended to the original speaker embedding extractor. The auxiliary network is designed to represent session information found within speaker embeddings. A key facet of our framework ensures that the primary speaker embedding extractor remains fixed during this process. Consequently, our system yields a twofold output; a similarity score reflecting speaker characteristics and another gauging session attributes. The latter, acting as a compensator, has the potential to rectify any discrepancies in the speaker score induced by analogous or differing session conditions. Our empirical evaluations, spanning various model architectures and evaluation configurations, underscore the feasibility of session compensation without the need for retraining the original embedding extractor.

In the field of speaker verification, various score calibration methods have been explored, including those using quality measure functions and embedding magnitude [10, 11]. Our proposed technique in this paper, however, stands out for two main reasons. First, it introduces a unique training method that deals solely with session information, allowing the use of existing embeddings without any modification. Second, this approach is designed to compensate for various environmental factors such as noise, channel differences, and recording conditions. Thus, our method offers easy integration with existing systems and the capability to adapt to various environments simultaneously.

The paper is organised as follows. Section 2 introduces the proposed framework. Experiments and result analysis are presented in Section 3, followed by conclusion in Section 4.
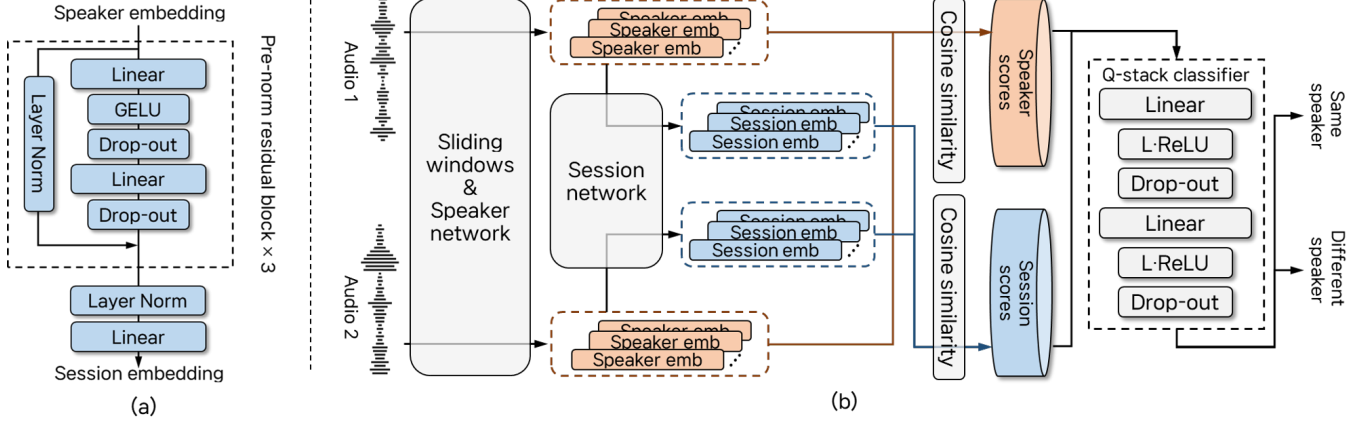
Fig. 1. The session compensating framework for speaker verification. (a) Illustration of the session network. The network receives speaker embeddings and, via multiple pre-norm residual blocks, produces session embeddings. (b) Outline of the speaker verification process within the proposed framework: Upon receiving two utterances, the system extracts corresponding session and speaker embeddings. Similarities between these embeddings are then calculated. The computed similarities are subsequently input into the Q-stack classifier to determine whether the two utterances originate from a same speaker or two distinct speakers.

2. FRAMEWORK FOR SESSION VARIABILITY COMPENSATION

In this section, we present a novel framework specifically designed to address and compensate for session variability in speaker verification tasks.

2.1. Speaker Embedding Extraction

For this study, we leverage pre-trained embedding extractors, drawing from methods that have proven efficacy in conventional recipes. Specifically, we have evaluated three models that represent a diverse cross-section of state-of-the-art architectures. These models are ECAPA-TDNN [12], RawNet3 [13], and MFA-Conformer-based speaker embedding extractors [14, 15].

2.2. Session Embedding Extraction

Within the domain of speaker verification, speaker embeddings efficiently capture the intrinsic attributes of a speaker's speech. However, these embeddings may also contain subtle information specific to the recording session, like background noise or recording device characteristics. Recognising the need to isolate such session-specific nuances from the core speaker features, we introduce the session network.

Network architecture. This network is attached to the speaker embedding network. Simplistically composed of several fully-connected layers, drop-out and GELU activation [16], the session network's primary role is to extract session information contained within the speaker embedding. Figure 1-(a) shows the detailed composition of the session network. It's designed to differentiate between the inherent speaker characteristics and the variabilities introduced by different recording sessions.

Training strategy. For effective extraction of session information,

it's paramount to train the network using a specially designed loss function. In addition, utilising datasets such as VoxCelebs [17, 18], which offers multiple sessions for individual speakers, is essential. For the session network, the training data comprises pairs – both positive and negative – drawn from the VoxCeleb datasets. These pairs are constructed by pairing two utterances. First, utterances for a positive pair stem from a same session and a same speaker, with identical augmentation techniques applied. This setup ensures that any discrepancy in the embeddings is predominantly due to session variations. Conversely, a negative pair includes two utterances from the same speaker but from different sessions, with distinct augmentations applied. This highlights the impact of session differences manifested within speaker embeddings. To elaborate further, consider a speaker denoted as i, randomly selected from our dataset. For our training, we aim to consider the speaker's utterances across two distinct sessions. Thus, for each chosen session, two random utterances are selected. This process gives us a notation, $u_{i,s,u}|s \in \{0,1\}, u \in \{0,1\}$, where i stands for the selected speaker, s denotes the session and u indicates the utterance. Now, for a definition of the loss function, we consider all possible combinations of sessions (s) and utterances (u). Our objective is to compute a loss value, $\mathcal{L}$, which would measure the difference or similarity between these combinations. This loss is determined as:

$$\mathcal{L} = \begin{cases} 1 - S(se(u_{i,s1,u1}), se(u_{i,s2,u2})), & \text{if } s1 == s2 \\ S(se(u_{i,s1,u1}), se(u_{i,s2,u2})), & \text{otherwise} \end{cases} \quad (1)$$

where $S(\cdot, \cdot)$ is a function indicating cosine similarity between two embeddings and $se(u)$ is session embedding from utterance u. It's worth noting that we do not consider pairs from different speakers while training the session network, ensuring the focus remains strictly on session variability. The session information is directly inferred from the video ID in the VoxCeleb datasets. In our context, two utterances are considered to be from the same session if they originate from an identical video.
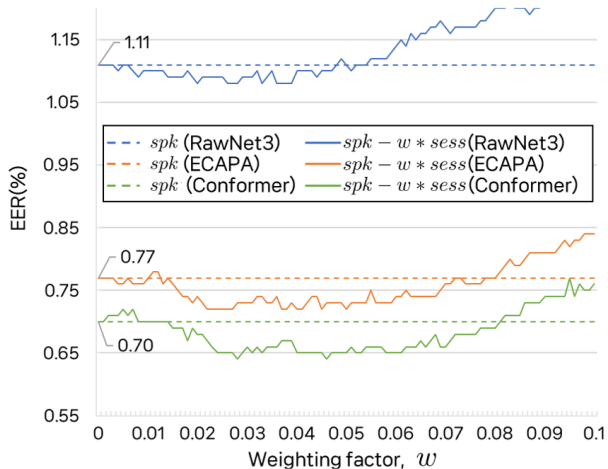
Fig. 2. Variation in speaker verification performance on the original VoxCeleb1 test set for three distinct embedding extractors. The graph shows the influence of session similarity($sess$)'s weight w on each extractor's performance. A clear trend emerges, highlighting the role of session similarity as a compensatory across all models evaluated.

2.3. Speaker Verification Using the Proposed Framework

In this section, we present our speaker verification procedure underpinned by our novel framework. In our study, we consider each verification trial to be constructed from a pair of utterances. From each of these utterances, two types of embeddings can be extracted: one that represents the characteristics of the speaker (the speaker embedding) and another that embodies the particularities of the recording session (the session embedding).

Score-level compensator. Once we have these embeddings, we can measure how similar they are. We compare the speaker embeddings from both utterances to get a "speaker similarity" score. This value essentially offers a metric that quantifies how alike the two utterances are, based on the characteristics of the speakers. On a parallel track, the session similarity is determined through the cosine similarity of the two session embeddings. This similarity shows how alike the two recordings are, based just on details from the recording session. Having obtained these similarities, the final step is to integrate them into a composite score that would be instrumental for verification. The formula we propose for this is:

$$score = spk - w * sess, \qquad (2)$$

where spk and $sess$ indicate speaker and session similarities, respectively, and w stands as a weighting factor for the session similarity. By subtracting a weighted session similarity from the speaker similarity, we aim to rectify any biases present in the speaker similarity attributed to session-related variations. Thus, the goal is to compensate for the session-induced biases, ensuring that the speaker's inherent characteristics shine through without the unexpected influence of session-specific attributes. To discern the impact of the session similarity on speaker verification, we carried out simple experiments utilising embeddings derived from the three embedding extractors. The focal point of this experiment was to adjust a weight value, and subsequently, observe how it influenced the performance of speaker verification. We conducted our tests using the VoxCeleb1 original test set, and the results are shown in Figure 2. The results reveal that simple action of subtracting the session similarity can reduce the error in speaker verification.

Q-stack-based compensator. Nonetheless, there exists a limitation to the above approach. The foundational premise of the approach is predicated on the assumption that the correlation between the speaker and session similarities is linear. However, in practical scenarios, this relationship might exhibit a more complex nature, suggesting the necessity for a sophisticated approach to accurately compensate for these interactions. To address this, we utilised an additional classifier which takes in both the speaker and session similarities and makes a binary decision. Essentially, it determines whether the two utterances originate from the same speaker or not. This new approach allows us to capture the non-linear relationship between the two similarities. The concept of this classifier is derived from a framework termed "Q-stack" [19]. The Q-stack classifier is employed to process two separate sets of similarities derived from two utterances, with the primary objective of deciding whether these utterances are from an identical speaker or not. The operation of the Q-stack-based framework is as follows. First, it takes in 200 similarities; half represents speaker similarities, and the other half stands for session similarities. These specific quantities originate from the well-known VoxCeleb trainer's recipe [1]. This procedure extracts 10 embeddings from an individual utterance through a sliding window technique. Consequently, when comparing a pair of utterances, the possible combination results in 10×10 similarities, leading to a combined total of 100 similarities for each type of embedding. For a more detailed architecture of the Q-stack, it is structured with three fully-connected layers, drop-out, and non-linear activation. These layers consist of 400 nodes, except the output layer with only two nodes. All hidden nodes are activated by leaky ReLU function for nonlinearity. Figure 1-(b) shows the overall operation of the proposed framework, including the structure of the Q-stack classifier.

3. EXPERIMENTS

3.1. Implementation details

For the evaluation of the proposed system, various datasets and models were employed. We selected multiple datasets for the training process: VoxCeleb1&2 [18, 20], VOiCES [21], CommonVoice [22] and telephone speeches from NIST SRE corpora. ECAPA-TDNN and RawNet3 models were trained using the VoxCeleb1&2 datasets. The Conformer-based system was trained leveraging the VoxCeleb1&2, NIST SRE 2004, 2006, and 2008 [23, 24], and CommonVoice datasets. The Q-stack

---

[1] https://github.com/clovaai/voxceleb_trainer

Table 1. A comparison of the performances using different models and evaluation sets. "Baseline" shows results from the usual speaker embedding. "Score comp" shows the outcomes when session variability is compensated at the score level. "Q-stack" denotes results when session variability is addressed using session embedding complemented by an additional classifier.

EER(%)	RawNet3				ECAPA-TDNN				Conformer			
	Vox1-O	N-SRE	VN-Mix	VC-Mix	Vox1-O	N-SRE	VN-Mix	VC-Mix	Vox1-O	N-SRE	VN-Mix	VC-Mix
Baseline	1.11	13.52	10.51	3.32	0.77	11.29	6.90	2.17	0.70	8.70	3.48	1.99
Score comp	1.12	13.33	8.91	3.05	0.75	10.92	5.84	2.02	0.69	8.58	3.43	1.88
Q-stack	1.06	12.98	7.34	3.03	0.71	10.64	4.22	1.98	0.65	8.39	3.34	1.51

Table 2. A comparison of the effect of the ensemble methods. "Single best" shows the top-performing model on its own. "Averaging scores" displays results when we combine scores from several models the usual way. "Proposed" gives results using our new ensemble method with Q-stack.

EER(%)	Vox1-O	N-SRE	VN-Mix	VC-Mix
Single best	0.70	8.70	3.48	1.99
Averaging scores	0.63	8.88	5.16	1.97
Proposed	0.56	8.14	3.17	1.44

system, distinctively, was trained on the test set of the VOiCES dataset. For augmentation, we use reverberations and noises from simulated RIRs and MUSAN datasets [25, 26]. Augmentation configurations follow that of [27].

3.2. Evaluation protocol

We evaluated performance using the VoxCeleb1 original test set (Vox1-O), 10sec-10sec protocol of NIST SRE 2010 evaluation (N-SRE) [28], and unique combined datasets. The initial evaluation of our system was carried out using two primary datasets: Vox1-O and N-SRE. These datasets contain audio data from varied sources and were chosen because they internally include session variability. To further evaluation, we introduced two custom datasets, VN-Mix and VC-Mix, crafted to test the systems' performance under challenging scenarios. First, VN-Mix (VoxCeleb and NIST) was composed of trials from Vox1-O and N-SRE. A notable aspect of this combination is the intrinsic domain difference between the two datasets. Specifically, N-SRE includes telephone speech while Vox1-O contains YouTube video clips. Given this contrast in source domains, it's hypothesised that a similarity bias might arise due to these inherent differences. For VC-Mix (VoxCeleb and VoxConverse), we combined positive pairs from Vox1-O with negative pairs from the "single" protocol of VoxConverse [29], as referenced in [30]. The positive pairs from Vox1-O comprise utterances from multiple sessions. In contrast, the negative pairs from VoxConverse are restricted to a singular session. This composition suggests the challenge, presenting both hard positive and negative pairs. In simple words, VC-Mix combines two types of pairs: one with the same speaker from different sessions and another with different speakers from a single session. The structure of VC-Mix was inspired by the dataset used in the VoxSRC 2022 challenge [31]. All telephone speech is up-sampled from 8kHz to 16kHz. The performance metric used to compare the models' performance was the well-known equal error rate (EER).

3.3. Comparison with single system

In Table 1, we presented a comprehensive comparison of the baseline system against our proposed systems across varied models and evaluation datasets. A key observation was the robustness and enhancement in EER offered by the proposed systems, which use session embeddings. Focusing on the "score comp" row, the results show the positive impact of session compensation using equation (2). The value of the weighting factor w was determined using test trials from the VOiCES dataset. Furthermore, the "Q-stack" row introduces further improvement from an additional classifier. This suggests that the classifier helps model a non-linear relationship between session and speaker similarities.

3.4. Comparison with ensemble system

Table 2 shows the impact of different ensemble techniques on model performance. A conventional ensemble averages multiple scores from various models. However, with our Q-stack system, this ensemble is more sophisticated. Instead of merely averaging, it inputs scores from different models in unison. In particular, we increased the number of input scores from 200 to 600 when combining the three models. The experimental results highlighted the superior performance of the Q-stack-based ensemble, especially on the N-SRE dataset and the VN-Mix containing the corresponding dataset. Conventional ensemble techniques, on the other hand, exhibited a decrement in performance on the N-SRE dataset, attributed to some models' limited exposure to telephone speech during their training.

4. CONCLUSION

In the domain of speaker verification, session variability is a well-known factor that can lead to performance degradation. Traditional methods often aim to modify or enhance the speaker embedding to handle this issue. Contrary to this, we suggest a novel approach; rather than adjusting the speaker embedding, we propose that session information should be treated as a separate entity. Comprehensive experiments, spanning a variety of models and datasets, demonstrate that the proposed method not only mitigates the effects of session variability but also has valuable implications for model ensemble and score calibration.

5. REFERENCES

[1] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[2] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*. IEEE, 2014, pp. 4052–4056.

[3] Jee-weon Jung, Hee-Soo Heo, Ju-ho Kim, Hye-jin Shim, and Ha-Jin Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *Proc. Interspeech 2019*, pp. 1268–1272, 2019.

[4] Weiwei Lin and Man-Wai Mak, "Robust speaker verification using deep weight space ensemble," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 802–812, 2023.

[5] Woo Hyun Kang, Sung Hwan Mun, Min Hyun Han, and Nam Soo Kim, "Disentangled speaker and nuisance attribute embedding for robust speaker verification," *IEEE Access*, vol. 8, pp. 141838–141849, 2020.

[6] Pavel Matějka, Ondřej Glembek, Fabio Castaldo, Md Jahangir Alam, Oldřich Plchot, Patrick Kenny, Lukáš Burget, and Jan Černocky, "Full-covariance ubm and heavy-tailed plda in i-vector speaker verification," in *Proc. ICASSP*. IEEE, 2011, pp. 4828–4831.

[7] Daniel Michelsanti and Zheng-Hua Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *Proc. Interspeech 2017*, pp. 2008–2012, 2017.

[8] Jaesung Huh, Hee Soo Heo, Jingu Kang, Shinji Watanabe, and Joon Son Chung, "Augmentation adversarial training for self-supervised speaker recognition," in *Workshop on Self-Supervised Learning for Speech and Audio Processing, NeurIPS*, 2020.

[9] Mohammad MohammadAmini, Driss Matrouf, Jean-François Bonastre, Sandipana Dowerah, Romain Serizel, and Denis Jouvet, "Learning noise robust resnet-based speaker embedding for speaker recognition," in *Odyssey 2022: The Speaker and Language Recognition Workshop*, 2022.

[10] Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck, "The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification," in *Proc. ICASSP*, 2021, pp. 5814–5818.

[11] Daniel Garcia-Romero, Gregory Sell, and Alan Mccree, "Magneto: X-vector magnitude estimation network plus offset for improved speaker recognition.," in *Odyssey*, 2020, pp. 1–8.

[12] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020, pp. 1–5.

[13] Jee-weon Jung, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung, "Pushing the limits of raw waveform speaker recognition," in *Proc. Interspeech*, 2022.

[14] Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung-yi Lee, and Helen Meng, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *Proc. Interspeech*, 2022.

[15] Jee-weon Jung, Hee-Soo Heo, Bong-Jin Lee, Jaesong Lee, Hye-jin Shim, Youngki Kwon, Joon Son Chung, and Shinji Watanabe, "Large-scale learning of generalised representations for speaker recognition," *arXiv preprint arXiv:2210.10985*, 2022.

[16] Dan Hendrycks and Kevin Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[17] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[18] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[19] Krzysztof Kryszczuk and Andrzej Drygajlo, "Improving classification with class-independent quality measures: Q-stack in face verification," in *Advances in Biometrics: International Conference, ICB 2007, Seoul, Korea, August 27-29, 2007. Proceedings*. Springer, 2007, pp. 1124–1133.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Proc. Interspeech*, 2017.

[21] Colleen Richey, Maria A. Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, Paul Gamble, Jeff Hetherly, Cory Stephenson, and Karl Ni, "Voices obscured in complex environmental settings (voices) corpus," 2018.

[22] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. LREC*, 2020.

[23] Mark A Przybocki, Alvin F Martin, and Audrey N Le, "Nist speaker recognition evaluations utilizing the mixer corpora—2004, 2005, 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1951–1959, 2007.

[24] Alvin F Martin and Craig S Greenberg, "Nist 2008 speaker recognition evaluation: Performance across telephone and room microphone channels," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[25] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*. 2017, pp. 5220–5224, IEEE.

[26] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv*, 10 2015.

[27] Hee Soo Heo, Bong-Jin Lee, Jaesung Huh, and Joon Son Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2009.14153*, 2020.

[28] Alvin F Martin and Craig S Greenberg, "The nist 2010 speaker recognition evaluation," in *Proc. Interspeech*, 2010.

[29] Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman, "Spot the conversation: Speaker diarisation in the wild," in *Proc. ICASSP*, 2020, pp. 299–303.

[30] Jee-weon Jung, Hee-Soo Heo, Bong-Jin Lee, Jaesung Huh, Andrew Brown, Youngki Kwon, Shinji Watanabe, and Joon Son Chung, "In search of strong embedding extractors for speaker diarisation," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[31] Jaesung Huh, Andrew Brown, Jee-weon Jung, Joon Son Chung, Arsha Nagrani, Daniel Garcia-Romero, and Andrew Zisserman, "Voxsrc 2022: The fourth voxceleb speaker recognition challenge," *arXiv preprint arXiv:2302.10248*, 2023.