INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

The VOiCES from a Distance Challenge 2019

Mahesh Kumar Nandwana[1], *Julien van Hout*[1], *Colleen Richey*[1], *Mitchell McLaren*[1], *Maria A. Barrios*[2], *Aaron Lawson*[1]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, California, USA
[2]Lab41, In-Q-Tel, Menlo Park, California, USA

{mahesh.nandwana, julien.vanhout, colleen.richey, mitchell.mclaren, aaron.lawson}@sri.com

Abstract

The VOiCES from a Distance Challenge 2019 was designed to foster research in the area of speaker recognition and automatic speech recognition (ASR) with a special focus on single-channel distant/far-field audio under various noisy conditions. The challenge was based on the recently released VOiCES corpus, with 60 international teams involved, of which 24 teams participated in the evaluation. In this paper, we separately present the challenge's speaker recognition and ASR tasks. For each task, we outline the training, development, and test data, as well as the evaluation metrics. Then, we report and discuss the results in light of the participant-provided system descriptions, to highlight the major factors contributing to high performance in distant speech processing.

Index Terms: VOiCES corpus, distant speech, speaker recognition, automatic speech recognition.

1. Introduction

SRI International, in collaboration with Lab41, organized "The VOiCES from a Distance Challenge 2019," focused on speaker and speech recognition on distant/far-field speech acquired using a single microphone in noisy and realistic reverberant environments [1]. The challenge was based on our recently released Voices Obscured in Complex Environmental Settings (VOiCES) corpus [2] and was held as a part of a special session at Interspeech 2019.

The main objectives of this challenge were to: i) benchmark state-of-the-art technology in the area of distant speaker recognition and automatic speech recognition (ASR); (ii) support the development of new ideas and technologies in speaker recognition and ASR; (iii) support new research groups entering the field of distant/far-field speech processing; and (iv) provide to the community a new, publicly available dataset that exhibits realistic distance characteristics.

The VOiCES corpus provides speech data recorded in acoustically challenging environments. Data was collected by recording retransmitted audio from high-quality loudspeakers in real rooms, capturing natural reverberation. LibriSpeech [3] was used as the clean speech source, while television, music, or babble played simultaneously from another loudspeaker as background noise. The speech loudspeaker rotated at predefined intervals during the recordings, to mimic human head movement. The dataset was released under the Creative Commons-BY 4.0 license, making it accessible for commercial, academic, and government use. More details on the VOiCES corpus can be found in [2, 4].

The VOiCES from a distance challenge consisted of two tasks: speaker recognition and ASR. Each task defined fixed and open system training conditions. Teams were allowed to participate in either the fixed condition, open condition, or both. These conditions were defined by the training data that could be used to train the systems. The fixed training condition served the purpose of benchmarking and comparing systems trained with the same data (or a subset thereof). The open training condition provided the means to quantify the gains that could be achieved with an unconstrained amount of data.

The challenge received tremendous response from the speech research community. A total of 60 international research organizations from academia and industry registered for this challenge. At the end of the evaluation phase, a total of 76 valid submissions from 24 teams were received, out of which 58 were submitted for the speaker recognition task, and 18 were submitted for the ASR task.

In the past, the CHiME speech separation challenge series [5], REVERB challenge [6], and IARPA Automatic Speech Recognition in Reverberant Environment (ASpIRE) challenge [7] have contributed to the growth of far-field ASR. But research in the speaker recognition community has tended to focus on close-talking data or data in the wild [8]. The fact that 21 teams participated in the speaker recognition task for the VOiCES 2019 challenge suggests a strong interest in freely available large scale datasets for benchmarking technology and driving research toward current and ongoing issues in the distant/far-field speech processing area.

In this work, we provide a summary of the evaluation submissions to draw attention to how current technology fairs on the VOiCES corpus. We anticipate that the results and publications that stem from the challenge and corresponding publicly available database will motivate the community to solve some of the remaining challenges in the far-field/distant speaker and speech recognition arena.

2. Speaker Recognition

In this section, we provide a brief overview of the training, development, and evaluation data for the speaker recognition task of the VOiCES challenge. The task of speaker recognition is: given a segment of speech and target speaker enrollment data, automatically determine whether the target speaker is speaking in the segment.

The speaker recognition task had fixed and open training conditions defined by the training data that could be used to train the system.

2.1. Training Data

In the *fixed* training condition, the system training data was restricted to specific datasets. Teams were only allowed to use the freely available Speakers in the Wild (SITW) [9], VoxCeleb1 [10], and VoxCeleb2 [11] datasets. The audio data from VoxCeleb1 and VoxCeleb2 was restricted to the official annotations for the fixed-condition submissions rather than allowing

http://dx.doi.org/10.21437/Interspeech.2019-1837

the use of video information from audio outside the annotations of each speaker. In addition, the teams were allowed to use publicly available non-speech audio and noises (e.g., noises, impulse responses, and compression technology) for data augmentation [12, 13], with the exception of MUSAN for babble, as it overlaps with the source data (LibriSpeech) of the VOiCES corpus. In contrast, the *open* training condition removed the limitations of the fixed condition and instead allowed the participants to use any proprietary and/or public data they had access to, including data from the fixed condition.

2.2. Development Data

The speaker recognition development set was created by subsetting the audio files from Rooms 1 and 2 (out of the four available) of the VOiCES corpus. The development set consisted of 15,904 audio segments from 196 speakers. Each audio file contained a single speaker. The development set represented different rooms, microphones, noise distractors, and loudspeaker angles. We ensured that the enroll and test audio segments corresponded to different book chapters from the source corpus (LibriSpeech) to prevent positive bias in performance from trials sourced from the same original session. The set included 20,224 target and 4,018,432 impostor trials. Table 1 provides a detailed breakdown of the development set across different parameters.

Table 1: *Details of the enrollment and verification sets for speaker recognition development across different parameters.*

	Enrollment Set	**Verification Set**
# Speakers	103	189
# Segments	256	15,648
Room ID	Room 1	Room 2
Mic. Type	Studio	Lapel
Mic. ID	01, 03	02, 04, 06, 08–12
Distractors	None	None, Music, TV, Babble
Loudspeaker Orientation	80°, 90°, 100°	0°, 60°, 90°, 120°, 180°

2.3. Evaluation Data

The speaker recognition evaluation set was created by subsetting the audio files of the VOiCES corpus to Rooms 3 and 4, as the associated source data. It consisted of 11,392 audio segments from 100 speakers that were disjoint from the development set. The evaluation set included 36,443 target and 357,1073 impostor trials. Table 2 details the breakdown of the evaluation set across different parameters.

Table 2: *Details of the enrollment and verification sets for speaker recognition evaluation across different parameters.*

	Enrollment Set	**Verification Set**
# Speakers	100	96
# Segments	326	11,066
Room ID	Room 3, Source	Room 4
Mic. Type	Studio	Lapel, Boundary, MEMS
Mic. ID	01, 03	04, 06, 08–12, 16–19
Distractors	None	None, TV, Babble
Loudspeaker Orientation	80°, 90°, 110°	0°, 30°, 60°, 90°, 120°, 150°, 180°

The evaluation set roughly mirrored the conditions of the development set; however, it contained three major sources of mismatch relative to the development set. First, two enrollment conditions included speakers who were enrolled by either using source data (i.e., no reverberation) or using data from Room 3, the latter being more similar to the development conditions. This element enabled assessing reverberation-mismatch impact. Second, the evaluation set included several unseen microphones relative to the development set. These new microphones were micro-electromechanical systems (MEMS) and boundary microphone types. This subset enabled analyzing the channel mismatch between the development and evaluation sets. Third, the rooms between the development and evaluation sets were disjoint to create an acoustic mismatch between sets. In particular, Room 3 was an "L" shaped room with very high reverberation characteristics.

2.4. Results

For the speaker recognition task, 21 teams successfully submitted their scores, out of which 4 teams submitted their scores for both the fixed and open conditions. We received 58 submissions, 50 of which were for the fixed training condition, and 8 were for the open training condition. The teams were allowed to submit up to three systems per condition. For the purpose of ranking teams, we picked the best score from each team as their official score for a given condition.

The primary metric for the speaker recognition task was a detection cost function based on the weighted sum of the miss and false-alarm error probabilities [14].

$$C_{det} = C_{miss} \times P_{miss} \times P_{tar} + C_{fa} \times P_{fa} \times (1 - P_{tar}) \quad (1)$$

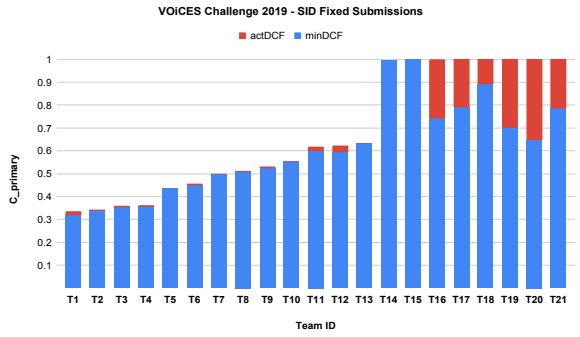
We assumed a prior target probability, P_{tar}, of 0.01 and equal cost of misses and false alarms. The C_{det} was normalized by the cost that a naïve system, which always chooses the least costly class, would get for the selected parameters. In our case, the normalization factor is given by P_{tar}.

Figure 1 represents the best actual and corresponding minimum $C_{primary}$ for the fixed and open conditions for all teams. It should be noted that a direct correlation exists between team IDs for the fixed and open condition results. For the fixed condition, it can be observed that the top teams exhibited highly impressive performance and that performance margin among the top four teams is very small. For the open training condition, the performance of the top two teams was quite competitive.

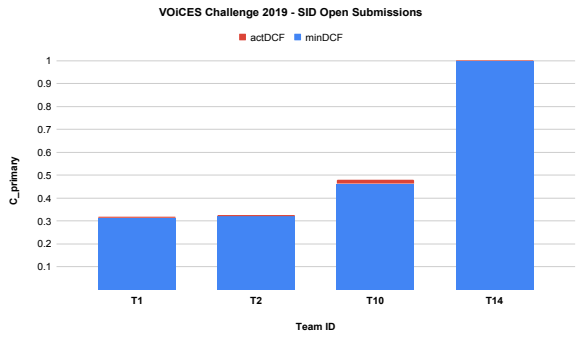
The major system components of the teams were speech activity detection (SAD); dereverberation; front-end feature extraction; embedding extractor; and probabilistic linear discriminant analysis (PLDA) back-end followed by score normalization, calibration, and multiple system fusion.

Most teams used energy-based SAD from the open-source KALDI toolkit [15]; while several teams used neural network based SAD. Some teams did not use speech activity detection at all, which given the speech-dense recordings may have seemed an appropriate choice. However, each of the top-performing teams applied SAD, implying its importance to the conditions. Some teams also applied dereverberation, with WPE (weighted prediction error) [16] being the most common method used prior to front-end feature extraction for speech dereverberation to improve signal quality.

The central theme across every system design was the use of the x-vector [12] architecture for speaker-embedding extraction. In the fusion step, x-vectors were complemented by modified versions of the x-vector achitecture and other architectures such as ResNet, ImageNet, and DenseNet. Some teams also



(a) *Fixed Condition*



(b) *Open Condition*

Figure 1: *Actual and minimum $C_{primary}$ for fixed and open speaker recognition submissions.*

used a traditional i-vector system [17] in their system fusion. For the training of the deep neural network embedding extractors, teams consistently augmented training data with different types of augmentation, such as noise, music, babbel, and reverberation [13, 18].

For back-end classification, a majority of the teams used PLDA. The teams also explored other back-end classifiers such as cosine distance for score computation and the top teams' used score normalization techniques, most commonly used techniques was adaptive S-norm.

System calibration was crucial in the challenge [19, 20] due to the acoustic mismatch between the dev and eval [21]. For the fixed condition, the top 13 teams achieved good calibration, with values of minimum C_{det} being very close to actual C_{det}. The six lowest-ranked teams did not successfully apply score calibration, which is reflected by the large difference between the minimum and actual costs (the red section of the bars in Figure 1a). The BOSARIS toolkit [22] was the most popular tool of choice for score calibration and system fusion.

3. Automatic Speech Recognition

In this section, we provide a brief overview of the training, development, and evaluation data used for the automatic speech recognition task of the VOiCES challenge. For benchmarking, the participants were expected to provide a transcript of each audio segment in a verbatim and case-insensitive manner.

The ASR task was evaluated over two training conditions: fixed and open. The two training conditions were defined by the specific datasets that could be used to build the ASR system. Teams could participate in either the fixed condition, open condition, or both.

3.1. Training Data

In the *fixed* condition, the training set consisted of an 80-hour subset of the LibriSpeech corpus. This subset was designed in such a way as to have no overlap in speakers with the VOiCES corpus (dev or eval). While the participants were allowed to train their own SAD as well as use external, non-speech resources for data augmentation (e.g., noises, impulse responses, and compression technology), they were not permitted to use additional speech data from any other source for model training (acoustic model, language model, speech enhancement, etc.). For data augmentation, teams were allowed to use babble from any publicly available resources, with the exception of MUSAN, as its babble data overlaps with the source data (LibriSpeech) of the VOiCES corpus.

In the *open* condition, participants could use any proprietary and/or public data they had access to along with the fixed-condition data. The only two datasets that participants were not allowed to use were: (1) any previously released subset of the VOiCES data and (2) any part of the LibriSpeech corpus outside of the 80-hour subset provided for the fixed condition.

3.2. Development Data

The ASR development set was disjoint from the speaker recognition development set and consisted of 20 hours of distant recordings from the 200 VOiCES dev speakers recorded in Rooms 1 and 2. As shown in Table 3, it contained recordings from 6 of the 12 mics, and was balanced across rooms, mics, distractor types, and loudspeaker angles. The metadata (mic, room, distractor, angle) were available in the filename, and the participants were allowed to use that information to analyze the behavior of their system under different conditions during development. The VOiCES Challenge development set was intended for teams to make system design decisions, but its use was not allowed during the direct training of SAD, enhancement processes, acoustic models or language models, in either the fixed or open conditions.

3.3. Evaluation Data

The ASR evaluation set was disjoint from the speaker recognition evaluation set and consisted of 20 hours of distant record-

Table 3: *Composition of the 20h ASR Development and Evaluation sets in terms of different rooms and microphones.*

microphone	mic-id	rm1	rm2	rm3	rm4
studio-close	1	Dev	Dev	Eval	-
lavalier-close	2	Dev	Dev	Eval	-
studio-middle	3	-	-	Eval	-
lavalier-middle	4	Eval	Eval	-	Eval
studio-far	5	Eval	Eval	-	Eval
lavalier-far	6	Dev	Dev	-	-
studio-behind	7	Eval	Eval	-	Eval
lavalier-behind	8	Dev	Dev	-	-
lavalier-under	9	Dev	Dev	-	-
lavalier-ceiling	10	Eval	Eval	-	Eval
lavalier-ceiling	11	Dev	Dev	-	-
barrier-table	16	-	-	Eval	Eval
mems-close	18	-	-	-	Eval
mems-fridge	19	-	-	-	Eval

ings from the 100 VOiCES eval speakers recorded in Rooms 1, 2, 3, and 4. As shown in Table 3, it contained recordings from 10 of the 20 mics. It was balanced across rooms, mics, distractor types, and loudspeaker angles. For Rooms 1 and 2, no microphone overlap exists between the dev and eval sets. Because of its size and shape, Room 3 is significantly harder for a given microphone, so we picked relatively easy mics. Room 4 is more similar to Rooms 1 and 2, so we picked different and/or more challenging microphones. Due its design, we expected teams to report higher errors on the evaluation set than on the development set.

3.4. Results

We received submissions from six teams in the ASR task. All six teams participated in the fixed condition, while only one team participated in the open condition.

We used the word error rate (WER) as the evaluation metric for the ASR portion of this challenge. NIST's open-source SCTK software was used to score participants' submissions by computing WER as the sum of errors (deletions, insertions, and substitutions) divided by the total number of words from the reference transcript. We used a GLM file to normalize potential differences in orthographic conventions. The results are shown

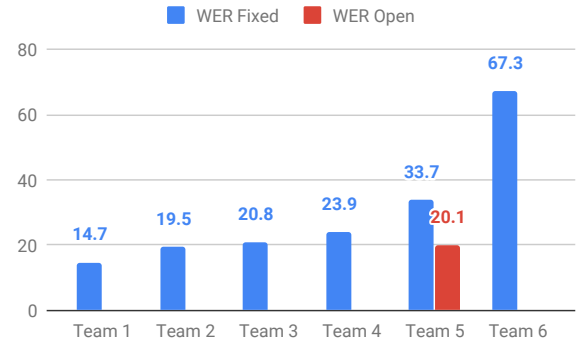


Figure 2: *ASR evaluation set word error rate (%).*

in Figure 2 with the team names anonymized. We report the results for both conditions together because only one team (Team 5) submitted outputs from systems trained in both conditions. For each condition, each team could submit up to three system outputs, and only the results from the best system per team were used in the rankings. Most teams chose to submit one or two outputs from a single system (without fusion), and the remaining outputs were from fusion of different systems. Only Team 6 chose to submit only single systems without any fusion.

The first observation is that the top teams achieved a decent WER in the range of 15–20% in conditions that are acoustically quite challenging. Although the amounts of training data available in the fixed condition were fairly limited, the teams' performance on the evaluation set varied widely, ranging from 14.7% to 67.3% WER. This points to the importance of careful system design when facing unseen and challenging acoustic conditions. In the next few paragraphs, we discuss some of the common or noteworthy design choices to shed some light on what approaches were most successful.

All teams used data augmentation to create artificially corrupted copies of the training data audio. Team 6 used a total of 480 hours of data; Team 3 and 4 used 960 hours; while Teams 1 and 5 used almost 2000 hours. All teams used simulations that included reverb as well as noises. For adding reverberation, Team 1 tried both time-domain [23] and frequency-domain [24] room simulations, and found frequency-domain to be superior; all other teams used time-domain added reverberation. Four of the teams (1, 2, 3, and 5) successfully used the weighted prediction error (WPE) algorithm [16] for de-reverberation. The top teams showed that many-fold data augmentation as well as WPE are critical to most effectively reduce the mismatch and distortions from distant speech.

Teams used a variety of neural network based acoustic models. Teams 1, 2, and 4 used the time-delay neural network (TDNN) architecture [25] in at least one of their subsystems; Teams 1, 3, and 5 used the more recent factorized-TDNN (F-TDNN) architecture [26]. Teams 1, 2, 3, and 5 also used convolutional architectures, and three teams (1, 2, and 5) reported using long-short term memory (LSTM) models [27]. Team 6 used acoustic models based on a light gated recurrent units (Li-GRU) architecture [28] implemented in the PyTorch-Kaldi toolkit [29]. The top four teams used speaker information in the form of i-vectors [17] (Teams 2, 3, 4) or x-vectors [12] (Team 1) to improve the acoustic models. Team 1 also trained novel embeddings targeted at predicting the room impulse response, called RIR-embeddings. Finally, Teams 2 and 5 used an attention mechanism [30] for some of their models, while Teams 1, 2, and 5 tried the recently introduced backstitch approach [31], although only Team 1 reported low WER gains of 0.3% from this approach. Based on the teams' system descriptions, the CNN-TDNN-F architecture appears the most popular and best performing on this task, especially with the addition of speaker and RiR-embeddings. None of the teams were successful in using end-to-end ASR models, probably due to the size of the training set in the fixed conditions.

For language modeling, the top five teams used LM rescoring using a neural network based LM. To tackle the out-of-vocabulary (OOV) problem, Team 1 trained a character-based LM and used it to generate artificial data with a much larger word coverage, thereby halving the OOV rate [32]. They combined the original and artificial text to train a final n-gram for decoding, obtaining a 1.1% absolute improvement over their already strong single-system baseline. Again, based on system descriptions by the top teams, lattice rescoring with a RNNLM was crucial to achieving the lowest possible word error rate.

To improve evaluation performance, the top five teams performed system fusion. While Team 2 performed fusion using the Recognizer Output Voting Error Reduction (ROVER) technique, the majority of the teams (1, 3, 4, and 5) performed fusion at the lattice level [33]. Finally, we point out that all six teams reported using the Kaldi ASR toolkit for at least some of their system building.

4. Conclusions

This paper presents a summary of the recently held "The VOiCES from a Distance Challenge 2019." The main goals of the challenge were to benchmark existing technology and support new ideas for distant/far-field speaker and speech recognition. We believe that this challenge and the VOiCES corpus will motivate advancement of research and technology development across reverberant and noisy conditions.

5. Acknowledgements

The authors would like to thank Dimitra Vergyri and Horacio Franco for the valuable discussions. We would also like to thank Chiachi Hung for his help with the system submission portal. This challenge was supported by SRI International.

6. References

[1] M. K. Nandwana, J. Van Hout, M. McLaren, C. Richey, A. Lawson, and M. A. Barrios, "The VOiCES from a distance challenge 2019 evaluation plan," *arXiv:1902.10828 [eess.AS]*, 2019.

[2] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, M. Graciarena, A. Lawson, M. K. Nandwana *et al.*, "Voices obscured in complex environmental settings (VOiCES) corpus," *Proc. Interspeech*, pp. 1566–1570, 2018.

[3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[4] M. K. Nandwana, J. van Hout, M. McLaren, A. Stauffer, C. Richey, A. Lawson, and M. Graciarena, "Robust speaker recognition from distant speech under real reverberant environments using speaker embeddings," *Proc. Interspeech*, pp. 1106–1110, 2018.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 504–511, 2015.

[6] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[7] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 547–554, 2015.

[8] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The 2016 speakers in the wild speaker recognition evaluation," *Proc. Interspeech*, pp. 823–827, 2016.

[9] ——, "The speakers in the wild (SITW) speaker recognition database." *Proc. Interspeech*, pp. 818–822, 2016.

[10] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech*, pp. 2616–2620, 2017.

[11] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech*, pp. 1086–1090, 2018.

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

[13] M. McLaren, D. Castan, M. K. Nandwana, L. Ferrer, and E. Yilmaz, "How to train your speaker embeddings extractor," *Speaker Odyssey*, pp. 327–334, 2018.

[14] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," *IEEE Signal Processing Society*, 2011.

[16] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[17] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[18] M. K. Nandwana, M. McLaren, D. Castan, J. van Hout, and A. Lawson, "Analysis of complementary information sources in the speaker embeddings framework," *Proc. Interspeech*, pp. 3568–3572, 2018.

[19] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[20] L. Ferrer, M. K. Nandwana, M. McLaren, D. Castan, and A. Lawson, "Toward fail-safe speaker recognition: Trial-Based Calibration with a reject option," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 140–153, 2019.

[21] M. K. Nandwana, L. Ferrer, M. McLaren, D. Castan, and A. Lawson, "Analysis of critical metadata factors for the calibration of speaker recognition system," *Proc. Interspeech*, 2019.

[22] N. Brümmer and E. De Villiers, "The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF," *arXiv preprint arXiv:1304.2865*, 2013.

[23] S. McGovern, "The image-source reverberation model in an n-dimensional space," *14th International Conference on Digital Audio Effects*, pp. 11–18, 2011.

[24] D. Campbell, K. Palomaki, and G. Brown, "A MATLAB simulation of" shoebox" room acoustics for use in research and teaching," *Computing and Information Systems*, vol. 9, no. 3, p. 48, 2005.

[25] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," *Proc. Interspeech*, pp. 3214–3218, 2015.

[26] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," *Proc. Interspeech*, pp. 3743–3747, 2018.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018.

[29] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," *arXiv preprint arXiv:1811.07453*, 2018.

[30] D. Povey, H. Hadian, P. Ghahremani, K. Li, and S. Khudanpur, "A time-restricted self-attention layer for ASR," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5874–5878, 2018.

[31] Y. Wang, V. Peddinti, H. Xu, X. Zhang, D. Povey, and S. Khudanpur, "Backstitch: Counteracting finite-sample bias via negative steps." *Proc. Interspeech*, pp. 1631–1635, 2017.

[32] Y. Y. Khokhlov, I. Medennikov, A. Romanenko, V. Mendelev, M. Korenevsky, A. Prudnikov, N. A. Tomashenko, and A. Zatvornitsky, "The STC keyword search system for OpenKWS 2016 evaluation." *Proc. Interspeech*, pp. 3602–3606, 2017.

[33] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.