ICASSP 2022 DEEP NOISE SUPPRESSION CHALLENGE

Harishchandra Dubey, Vishak Gopal, Ross Cutler, Ashkan Aazami, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, Robert Aichner

Microsoft Corporation, Redmond, USA
firstname.lastname@microsoft.com

ABSTRACT

The Deep Noise Suppression (DNS) challenge is designed to foster innovation in the area of noise suppression to achieve superior perceptual speech quality. This is the 4th DNS challenge, with the previous editions held at INTERSPEECH 2020 [1], ICASSP 2021 [2], and INTERSPEECH 2021 [3]. We open-source datasets and test sets for researchers to train their deep noise suppression models, as well as a subjective evaluation framework based on ITU-T P.835 to rate and rank-order the challenge entries. We provide access to DNSMOS P.835 and word accuracy (WAcc) APIs to challenge participants to help with iterative model improvements. In this challenge, we introduced the following changes: (i) Included mobile device scenarios in the blind test set; (ii) Included a personalized noise suppression track with baseline; (iii) Added WAcc as an objective metric; (iv) Included DNSMOS P.835; (v) Made the training datasets and test sets fullband (48 kHz). We use an average of WAcc and subjective scores P.835 SIG, BAK, and OVRL to get the final score for ranking the DNS models. We believe that as a research community, we still have a long way to go in achieving excellent speech quality in challenging noisy real-world scenarios.

Index Terms— Deep Noise Suppression, P.835, Perceptual Speech Quality, Personalized Noise Suppression, Speech Enhancement

1. INTRODUCTION

In recent times, hybrid work has become the "new normal" as the number of people working remotely has increased significantly due to the COVID-19 endemic. Audio calls in the presence of background noises such as a dog barking, a baby crying, kitchen noises, neighboring talkers, in-car noises, etc., get significantly degraded in terms of quality/intelligibility of the perceived speech. This leads to increased fatigue in audio meetings. Deep learning-based noise suppression (DNS) has shown promising results with superior speech quality [4, 5, 2] which is significantly better than classical approaches [6]. Previous DNS challenges accelerated DNS research by providing a massive training dataset, real test sets, training data synthesizer, and subjective evaluation frameworks based on ITU-T P.808 [7], and P.835 [8]. Many recent papers have leveraged the DNS challenge datasets for developing DNS models [1, 2, 3].

The ICASSP 2022 DNS challenge focuses on personalized and non-personalized DNS for fullband audio. Personalized denoising suppresses neighboring talkers in addition to background noises. We provide fullband datasets for training personalized and non-personalized DNS models. We collected real-world test sets, developed a framework for P.835 subjective evaluation, created APIs for DNSMOS P.835 [9] and WAcc. We provided the development test set, DNSMOS P.835 API, and WAcc API at the beginning of the challenge, which helps participants to optimize their models. The blind test set is released 5 days before the challenge deadline, and this is used for ranking the models for DNS performance. We evaluated the submitted models based on ITU-T P.835 subjective evaluation scores, namely speech quality (SIG), background noise quality (BAK), and overall audio quality (OVRL), as well as WAcc from a state-of-the-art speech recognition system. In addition, we make DNSMOS P.835 [9] freely available to researchers. DNSMOS P.835 is a deep learning model that predicts SIG, BAK, OVRL scores for a noisy test clip.

2. CHALLENGE TRACKS

This challenge has two tracks, namely (1) non-personalized DNS and (2) personalized DNS (PDNS) for fullband audio. Unlike previous challenges, this time, we did not have wideband (16 kHz) data in our training data and testset. Similar to previous DNS Challenges, we provide a training data synthesizer that could also be used with other datasets if participants choose to do so. The data synthesizer, configuration, scripts to access the Azure APIs, and data download scripts are provided in the challenge Github repository[1].

We provided a baseline model for Track 1[2] and baseline enhanced test clips for Track 2. In this paper, we briefly describe the baseline models for both tracks. We adopted WAcc as an objective metric for measuring the impact of DNS on speech recognition systems. Participants submitted enhanced clips for one or both tracks. We conducted ITU-T P.835 and WAcc computation on submitted enhanced clips. The motivation to add WAcc as an evaluation metric stems from the fact that several models from past challenges had noticeable WAcc degradation resulting from over-suppression of noise and/or speech distortions. We provided the participants with an Azure API for estimating WAcc on the development set. We computed DNSMOS P.835 [9] for each audio clip in the training set and provided this to participants. DNSMOS P.835 scores can be used to segment the training dataset for conducting the data ablation studies. Participants can do experiments with different portions of the training dataset based on a chosen threshold for SIG, BAK, and OVRL. The computational requirements for the challenge tracks are described in https://aka.ms/dns-challenge

3. DATASETS

3.1. Training Datasets

We provide clean speech, noise, impulse responses, and a training data synthesizer for both tracks. The same noise and impulse responses are provided for both tracks. Each track has its training data synthesizer. Our training data consists of English read speech,

[1] https://github.com/microsoft/DNS-Challenge

[2] https://github.com/microsoft/DNS-Challenge/blob/master/NSNet2-baseline/nsnet2-20ms-48k-baseline.onnx

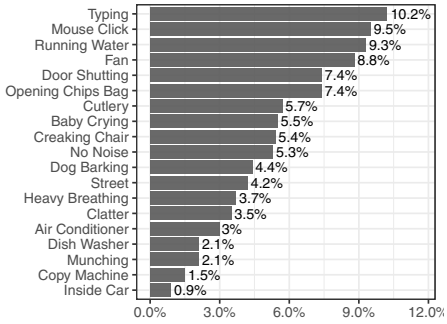

Fig. 1. Distribution of noise types in our blind test set.

English singing, French, German, Italian, Russian, and Spanish languages. The PDNS track has clean speech, where each audio clip consists of a concatenation of all audio clips belonging to a talker. We also provide a baseline speaker embedding for each talker in the PDNS training set. We choose clean speech with DNSMOS P.835 SIG, BAK, and OVRL to have a score of more than 4.25. The PDNS track leverages clean speech with a DNSMOS P.835 OVRL $\geq$ 4.25. Next, the audio files for each talker are combined into a single clip. We randomly sample 2.5 minutes of clean speech for each speaker and provide it as enrollment speech. We extract speaker embeddings for each enrollment audio using a baseline RawNet2 speaker model [10] which is adopted as the baseline speaker embedding extractor for this challenge. The next sections describe the clean speech and noise data.

3.1.1. Clean Speech

The clean speech consists of six languages, namely English, French, German, Italian, Russian, and Spanish. English clean speech consists of read speech and singing, while the rest of the languages only have read speech. We provide DNSMOS P.835 scores to help participants filter the data based on DNSMOS P.835 scores if they want to do data ablation studies. The personalized track consists of clips with DNSMOS P.835 OVRL $\geq$ 4.25. We combine all audio clips from each unique talker into a single file. The PDNS training data has a total of 3230 talkers, out of which 60% of the talkers are randomly chosen to be primary talkers while the rest are neighboring talkers. We provide the file list for PDNS clean speech with 'primary' and 'secondary' tags. Challenge participants can use the provided primary/secondary tags or generate their own. We sampled 60% of the speakers as primary talkers, ensuring a uniform distribution among all languages, read speech or singing.

English clean speech is derived from Librivox[3] where we include audio clips chosen using a subjective ITU-T P.808 framework [7]. English singing data consists of high-quality audio recordings from professional singers contained in the *VocalSet* corpus [11]. It has 10.1 hours of clean singing recorded by 20 professional singers: 9 males, and 11 females. This data was recorded on a range of vowels, a diverse set of voices on several standard and extended vocal techniques, and sung in contexts of scales, arpeggios, long tones, and excerpts. The PDNS English clean speech contains 1934 talkers from Librivox, 110 talkers from VCTK corpus, 20 talkers from Vocalset. The PDNS clean speech consists of 47 talkers for French, 874 for German, 14 for Italian, 7 for Russian, and 224 for Spanish.

[3]https://librivox.org/

3.1.2. Noise

We use the same noise clips for both tracks. Noise data consists of about 62,000 clips belonging to 150 noise classes. The noise clips were chosen from Audio Set[4] [12] and Freesound[5]. Audio Set is a collection of about 2 million human-labeled 10s sound clips drawn from YouTube videos belonging to about 600 audio events. Certain audio event classes are over-represented in Audio Set. For example, there are over a million clips with audio classes music and speech and less than 200 clips for classes such as toothbrush, creaking, etc. Approximately 42% of the clips have a single class, but the rest may have 2 to 15 labels. We developed a sampling approach to balance the noise classes in our dataset such that each class has at least 500 clips. We used a speech activity detector to remove the clips with any kind of speech activity (voice content) from our noise clips. This enabled us to get noise clips with no presence of speech. We augmented the Audio Set noise clips with 10,000 noise clips downloaded from Freesound and DEMAND databases [13]. The total noise data constitute 181 hours of audio. Fig. 1 shows the histogram of noise classes included in the blind test set. These noise types were validated by a human listener after the collection of the blind set.

3.1.3. Impulse Responses

We provide 248 real and about 60,000 synthetic room impulse responses, which can be leveraged for generating reverberant noisy training data. The training data synthesizer adds noise to reverberant clean speech depending on the chosen configuration. Participants may choose to use clean speech or reverberant speech as training targets for their DNS models. We chose impulse responses from the openSLR26 and openSLR28 [14] datasets.

3.2. Test set

We have two test sets: dev test set and blind test set. The dev test set is intended for model development and optimization, and is provided at the start of the challenge. The blind test set is used for ranking the challenge model in terms of evaluation metrics and is intended to be used as an unseen test set. Good performance of a model on the blind test set will show it is generalized. Both test sets consist of fullband audio clips recorded in real-world scenarios collected through crowd-sourcing where workers read provided text prompts and record their voice using desktop/laptop/mobile devices in the presence of noise and/or neighboring talkers.

3.2.1. Non-personalized Development Test set

The development test set for the non-personalized track consists of 930 real recordings. All clips contain noisy speech in the English language. Among these, 193 test clips have emotional speech in the presence of noise. There are six emotion types, namely happy, sad, angry, yelling, crying, and laughter. Crowd-sourced workers were asked to read provided text prompts and create emotional events in each test clip. The remaining clips contain the voice of a talker reading text in the presence of the following noise types: fan, air conditioner, typing, door shutting, clatter noise, car noise (i.e., standing near a car on a busy street or standing outside the car), kitchen noise (noise from kitchen utensils, dish scrubbing etc.), dish washer, running water, opening chips bags, munching or eating, creaking chair, heavy breathing, copy machine, baby crying, dog barking, inside-car noise (e.g., sitting on a passenger seat in a car which is being driven by someone else), mouse clicks, mouse scroll wheel, touch

[4]https://research.google.com/AudioSet
[5]https://freesound.org/

pad clicks, etc. Each test clip was recorded at 48 kHz with a duration of 10 to 20 seconds. Workers were asked to record in the near-field (close-talk) and far-field with distances of 1, 2, and 3 meters. All test clips in the non-personalized development test set were recorded using a laptop or desktop computer.

3.2.2. Personalized Development Test set

Both development and blind test set have 2.5 minutes of enrollment speech for primary talkers to be used in personalized denoising. PDNS leverages speaker embedding (features) for preserving only the primary talker in a noisy environment while suppressing the neighboring talkers and noise. The development test set for the personalized track consists of 1443 real recordings. All clips contain noisy speech in the English language. Among these, 193 test clips have emotional speech in the presence of noise and are identical to the emotional test clips in the non-personalized track. There are 737 test clips where the primary talker reads the provided text in the presence of the same noise types as those in the non-personalized track. Each test clip was recorded at 48 kHz with a duration of 10–20 seconds. Workers were asked to record in the near-field (close-talk) and far-field with distances of 1, 2, and 3 meters. There are 166 test clips with the primary talker speaking in the presence of a neighboring talker and noise where both the noise and neighboring talker are simultaneously active in the primary talker's background. There are 347 test clips where the primary talker is speaking in the presence of a neighboring speaker with no background noise. Thus, we have simulated three scenarios for PDNS: (i) primary talker in the presence of noise; (ii) primary talker in the presence of neighboring talker; and (iii) primary talker in the presence of simultaneously active neighboring talker and noise. All test clips in the personalized development test set were recorded using a laptop or desktop computer.

3.3. Blind test set

We collected a common blind test set for both tracks, which facilitates a direct comparison of both tracks to elucidate the benefits of personalized noise suppression. Track 2 leverages 2.5 minutes of clean speech enrollment data for personalized denoising. The blind set has 2.5 minutes of enrollment speech for each test clip, which is intended for use only by personalized models. There are 859 real test clips, each with a duration of 10s, in the blind test set. We collected the blind test set on a variety of desktop and mobile platforms through crowd-sourcing, where 70% of the test clips were collected on a smartphone. The blind test set went through several iterations of data validation based on unit tests and human listening. We did not include test clips from the same speaker if they had a similar background noise. In this way, we have a unique speaker and background noises in each test clip. We transcribed the blind test set using a third-party data annotation service. We did expert listening to correct the speech transcription for the blind test set to ensure high accuracy. Participants were provided with the blind test set 5 days before the data submission deadline.

4. RESULTS & DISCUSSIONS

4.1. Baseline for Non-personalized DNS

We trained NSNet2 [3, 15] on a subset of the fullband non-personalized training dataset to obtain the baseline. The subset was characterized by clean speech clips having the DNSMOS P.835 scores greater than or equal to 4.2, 4.5, and 4.0 for SIG, BAK, and OVRL, respectively. We denoised the Track 1 dev test-set and blind test-set with the trained NSNet2 baseline model, to get the baseline enhanced clips. We release the ONNX[6] model and inference script for NSNet2.

4.2. Baseline for Personalized DNS

The baseline for the PDNS track consists of the baseline speaker embedding extractor and the baseline PDNS model. The baseline speaker embedding extractor is RawNet2 [16] trained on wideband audio from VoxCeleb2 [17]. Lack of fullband audio datasets with speaker IDs and thousands of speakers led us to select VoxCeleb2, which is a wideband audio dataset with 6,000 speakers. Thus our baseline PDNS model had a wideband speaker embedding extractor. We used the baseline speaker embedding extractor on all enrollment data in the training set, and provided the participants with speaker embeddings. This lowered the barrier to entry in the PDNS track. Participants were permitted to retrain the RawNet2 model with fullband data. The challenge permitted the use of external datasets in addition to DNS challenge datasets. Participants could leverage other state-of-the-art speaker models [18, 19, 16] for extracting speaker embeddings [18, 19].

PDNS models are trained to suppress neighboring talkers and background noises and only preserve the enhanced speech from the primary talker. To achieve this, PDNS models leverage speaker embedding features extracted from the enrollment speech along with spectral features (or raw-waveform) of the noisy input audio. We used the personalized DCCRN (pDCCRN) model described in [20] as the baseline PDNS model for this challenge. We modified pDCCRN to accept 48 kHz input and added a layer to the encoder and decoder. Since the baseline speaker embedding extractor uses wideband audio, the input audio is downsampled to 16 kHz before extracting the speaker embeddings.

Each unique talker in the personalized training dataset, PDNS dev test set, and blind test set has 2.5 minutes of enrollment speech. PDNS models are expected to leverage talker-aware training and talker-adapted inference. There are two motivations to provide clean speech for enrollment of the primary talker: (1) speaker models are sensitive to false-alarms in speech activity detection (SAD) [21]; clean speech can be used for obtaining accurate SAD labels resulting in speaker-discriminative embeddings. (2) Speaker adaptation is expected to work well using multi-conditioned data; clean speech can be used for generating reverberant and noisy multi-condition enrollment data for speaker adaptation.

4.3. Evaluation Methodology

This challenge relies on ITU-T P.835 [8] subjective evaluation for evaluating the DNS models, since objective speech quality metrics, such as PESQ [22], SDR, and POLQA [23], do not correlate well with subjective speech quality [24]. A modified version of ITU-T P.835 was used for measuring the performance of personalized models. The modified P.835 for personalized DNS provides 10s of enrollment speech of the primary speaker so that the raters can recognize the primary talker's voice while assigning subjective scores. Human raters were instructed to focus on the quality of the voice of the primary talker when more than two talkers were present in a test clip.

Four metrics, three P.835 subjective scores (SIG, BAK, OVRL) and WAcc from a speech recognition system were used to evaluate challenge models. Scores on the blind test set were combined into a final score for ranking the models. Higher WAcc shows superior denoising performance with respect to speech recognition. WAcc is defined as WAcc $= 1 -$ WER, where WER is the word error rate of

[6] https://onnx.ai

Team#	dSIG	dBAK	dOVRL	dWacc	Final Score
2	0.00	2.54	1.49	-0.02	0.74
14	-0.03	2.12	1.25	-0.03	0.71
19	-0.10	2.11	1.23	-0.04	0.70
41	-0.19	2.30	1.22	-0.04	0.69
25	-0.27	2.39	1.19	-0.07	0.68
46	-0.27	2.08	1.06	-0.05	0.67
15	-0.25	2.27	1.13	-0.07	0.67
45	-0.28	2.06	1.03	-0.05	0.67
29	-0.32	2.10	0.99	-0.04	0.67
3	-0.32	2.26	1.09	-0.07	0.67
7	-0.30	2.03	0.98	-0.04	0.67
11	-0.43	2.31	1.03	-0.07	0.66
37	-0.32	2.06	0.97	-0.07	0.65
22	-0.17	1.50	0.83	-0.05	0.64
43	-0.35	1.70	0.83	-0.05	0.64
33	-0.31	1.62	0.84	-0.06	0.64
47	-0.46	1.70	0.73	-0.05	0.63
16	-0.58	2.07	0.84	-0.10	0.62
54	-0.57	1.87	0.74	-0.08	0.62
35	-0.55	1.92	0.75	-0.09	0.61
49	-0.68	1.94	0.70	-0.10	0.60
Baseline	-0.66	1.78	0.63	-0.09	0.60
39	-0.35	1.15	0.53	-0.08	0.59
52	-0.39	0.89	0.37	-0.04	0.59
Noisy	0.00	0.00	0.00	0.00	0.56
31	-0.26	1.56	0.80	-0.70	0.31
51	-2.24	1.44	-0.73	-0.70	0.12

Fig. 2. Results: P.835 subjective evaluation of all models from Track 1 non-personalized DNS on blind testset.

Team#	dSIG	dBAK	dOVRL	dWAcc	Final Score
17	-0.05	2.36	1.40	-0.02	0.72
42	-0.06	2.40	1.41	-0.03	0.72
19	-0.08	2.14	1.27	-0.03	0.70
29	-0.36	2.18	1.07	-0.04	0.67
31	-0.26	1.60	0.86	-0.05	0.64
15	-0.52	2.34	1.00	-0.11	0.63
Baseline	-0.61	2.09	0.84	-0.08	0.62
44	-0.69	2.11	0.80	-0.13	0.59
49	-0.74	1.73	0.60	-0.09	0.59
6	-0.51	1.24	0.53	-0.10	0.57
Noisy	0.00	0.00	0.00	0.00	0.56
13	-1.11	1.29	0.08	-0.23	0.45

Fig. 3. Results: P.835 subjective evaluation of all models from Track 2 personalized DNS on blind testset.

speech recognition system. The final score is computed as Final score $= 0.5[\text{WAcc} + 0.25(\text{OVRL} - 1)]$.

4.4. Results

We received 24 and 10 submissions for Track 1 and Track 2, respectively. Each team submitted a processed blind test set (Sec. 3.3).

Fig. 2 and Fig. 3 show the subjective P.835 scores, WAcc and final score for challenge entries in decreasing order of performance. dSIG, dBAK, dOVRL refers to the difference in SIG, BAK, OVRL between the enhanced clip and noisy clip. Similarly, dWAcc is the difference in WAcc between the enhanced clip and noisy clips.

For the top performing teams, we ran an ANOVA test to determine statistical significance (see https://aka.ms/dns-challenge). The 2nd, 3rd and 4th place are tied for Track 1. Likewise the 1st and 2nd place for Track 2 are tied. Teams 17, 19, and 42 did not submit a paper so were disqualified per the challenge rules.

From a breakdown of scores based on the device type (mobile/desktop) we find that the MOS scores for clips recorded on mobile devices is higher than those from desktop devices (see https://aka.ms/dns-challenge). This suggests that mobile had better acoustic devices or environments than the desktop scenarios.

We required participants to not do any automatic gain control (AGC). Also, we did not perform any AGC on blind test clips or enhanced clips. A state-of-the-art speech recognition API from Azure Cognitive service was used for computing WAcc. The speech recognition system was trained to handle audio with a wide range of energy levels so we do not expect any degradation of WAcc due to varying energy levels in clips from the blind test set. Table 1 shows the Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRCC) between per-model subjective scores and corresponding DNSMOS P.835 scores [9]. The high correlation between subjective scores and DNSMOS P.835 in both tracks shows the efficacy of DNSMOS P.835 in ranking the DNS models. It validated our approach for providing a dev test set and DNSMOS P.835 to challenge participants for model development.

Table 2 gives a high-level comparison of the top performing models. Note there is low correlation of model size or real-time factor with performance. The top performing models for Track 1 didn't use additional datasets, while the Track 2 models did. The winning team for Track 1 [25] also won the ICASSP 2022 AEC Challenge [26], and demonstrates a single model can provide excellent AEC, DNS, and WAcc performance. The performance of the personalized DNS track also show excellent performance, greatly exceeding results of our first personalized DNS challenge [2] with the winner [27] providing very good dOVRL with low dSIG and low dWAcc. Note, however, that no team actually improves SIG and WAcc is 2% worse than noisy. There is still a lot of room for improvement.

Table 1. DNSMOS PCC and SRCC

	Track 1			Track 2		
	SIG	BAK	OVRL	SIG	BAK	OVRL
PCC	0.93	0.92	0.94	0.92	0.96	0.96
SRCC	0.78	0.89	0.85	0.84	0.89	0.93

Table 2. Comparison of top performing models.

Track	Team	Params	Real-time Factor	Additional data-sets
1	2 [28]	1.5M	0.60	N
1	14 [29]	10.27 M	0.68	N
1	41 [30]	29.9 M	0.45	N
1	25 [31]	5.29 M	0.65	N
2	42 [27]	7.81 M	0.96	Y
2	29 [32]	12.41 M	0.19	Y

5. CONCLUSION

We hope this challenge dataset, test set, test framework, DNSMOS P.835, and top performing papers (Table 2) help push the field forward. The next DNS challenge will have a more diverse test set including more languages, accents, and devices from realistic noisy scenarios. We plan to have a dedicated inference engine/evaluation setup for computing the model complexity and inference time for all submitted models. We will also include a validation of the look-ahead to make sure the comparison is fair for all models.

6. REFERENCES

[1] C. K. Reddy et al., "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *ISCA INTERSPEECH*, 2020.

[2] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 Deep Noise Suppression Challenge," in *IEEE ICASSP*, 2021, pp. 6623–6627.

[3] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 Deep Noise Suppression Challenge," *ISCA INTERSPEECH*, 2021.

[4] H.-S. Choi, H. Heo, J. H. Lee, and K. Lee, "Phase-aware single-stage speech denoising and dereverberation with U-net," *arXiv preprint arXiv:2006.00687*, 2020.

[5] Y. Koyama, T. Vuong, S. Uhlich, and B. Raj, "Exploring the best loss function for DNN-based low-latency speech enhancement with temporal convolutional networks," *arXiv preprint arXiv:2005.11611*, 2020.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE TASLP*, 1984.

[7] B. Naderi and R. Cutler, "An open source implementation of ITU-T recommendation P.808 with validation," in *ISCA INTERSPEECH*, 2020.

[8] B. Naderi and R. Cutler, "Subjective evaluation of noise suppression algorithms in crowdsourcing," in *ISCA INTERSPEECH*, 2021.

[9] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE ICASSP*, 2021.

[10] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *ISCA INTERSPEECH*, 2019.

[11] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset.," in *ISMIR*, 2018.

[12] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE ICASSP*, 2017.

[13] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, p. 3591, 05 2013.

[14] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE ICASSP*, 2017.

[15] S. Braun and I. Tashev, "Data augmentation and loss normalization for deep noise suppression," in *International Conference on Speech and Computer*. Springer, 2020, pp. 79–86.

[16] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," *ISCA INTERSPEECH*, 2020.

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *ISCA INTERSPEECH*, 2018.

[18] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *IEEE ICASSP*, 2018, pp. 4879–4883.

[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE ICASSP*, 2018, pp. 5329–5333.

[20] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized speech enhancement: New models and comprehensive evaluation," *arXiv preprint arXiv:2110.09625*, 2021.

[21] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[22] "ITU-T recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb 2001.

[23] J. Beerends et al., "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part II-perceptual model," *AES: Journal of the Audio Engineering Society*, vol. 61, pp. 385–402, 06 2013.

[24] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *IEEE ICASSP*, 2019.

[25] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," *ICASSP*, 2022.

[26] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, and R. Aichner, "ICASSP 2022 Acoustic Echo Cancellation Challenge," in *ICASSP*, 2022.

[27] Y. Ju, W. Rao, X. Yan, Y. Fu, S. Lv, L. Cheng, Y. Wang, L. Xie, and S. Shang, "TEA-PSE: Tencent-ethereal-audio-lab personalized speech enhancement system for ICASSP 2022 DNS CHALLENGE," in *IEEE ICASSP*, 2022.

[28] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *IEEE ICASSP*, 2022.

[29] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *IEEE ICASSP*, 2022.

[30] Z. Zhang, L. Zhang, X. Zhuang, Y. Qian, H. Li, and M. Wang, "FB-MSTCN: A full-band single-channel speech enhancement method based on multi-scale temporal convolutional network," in *IEEE ICASSP*, 2022.

[31] T. Wang, W. Zhu, Y. Gao, Y. Chen, J. Feng, and S. Zhang, "Harmonic gated compensation network plus for ICASSP 2022 dns challenge," in *IEEE ICASSP*, 2022.

[32] L. Chen, C. Xu, X. Zhang, X. Ren, X. Zheng, C. Zhang, L. Guo, and B. Yu, "Multi-stage and multi-loss training for fullband non-personalized and personalized speech enhancement," in *IEEE ICASSP*, 2022.