EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation

Julius Richter[1], *Yi-Chiao Wu*[2], *Steven Krenn*[2], *Simon Welker*[1], *Bunlong Lay*[1], *Shinji Watanabe*[3], *Alexander Richard*[2], *Timo Gerkmann*[1]

[1]Signal Processing (SP), Universität Hamburg, Germany
[2]Codec Avatars Lab, Meta, Pittsburgh, PA, USA
[3]Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA, USA

julius.richter@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

Abstract

We release the EARS (**E**xpressive **A**nechoic **R**ecordings of **S**peech) dataset, a high-quality speech dataset comprising 107 speakers from diverse backgrounds, totaling in 100 hours of clean, anechoic speech data. The dataset covers a large range of different speaking styles, including emotional speech, different reading styles, non-verbal sounds, and conversational freeform speech. We benchmark various methods for speech enhancement and dereverberation on the dataset and evaluate their performance through a set of instrumental metrics. In addition, we conduct a listening test with 20 participants for the speech enhancement task, where a generative method is preferred. We introduce a blind test set that allows for automatic online evaluation of uploaded data. Dataset download links and automatic evaluation server can be found online[1].

Index Terms: speech dataset, speech enhancement, dereverberation, benchmark

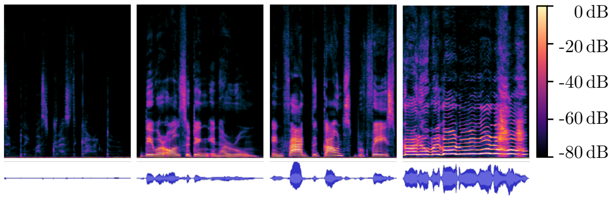
1. Introduction

Learning-based speech processing has seen huge leaps forward in recent years, with the impact of deep learning spanning essentially all areas from speech representation learning [1] over text-to-speech [2] to speech enhancement [3]. Publicly available datasets such as LibriSpeech [4] or VCTK [5] have undoubtedly been a key driver of open and reproducible research in our field and have enabled steady progress. However, these datasets typically come with multiple shortcomings and are either too small, of low recording quality or do not span a large enough variety of different speakers and speaking styles.

To overcome these shortcomings, we release the **E**xpressive **A**nechoic **R**ecordings of **S**peech (EARS) dataset. EARS contains 100 h of anechoic speech recordings at 48 kHz from over 100 English speakers with high demographic diversity. The dataset spans the full range of human speech, including reading tasks in seven different reading styles, emotional reading and freeform speech in 22 different emotions, conversational speech, and non-verbal sounds like laughter or coughing.

In addition, we set up a speech enhancement and speech dereverberation benchmark on EARS, comparing several predictive [6, 7] and generative [8, 9] speech enhancement methods. The benchmarks are intended to provide valuable insights into models' strengths, limitations, and comparability, thus promoting the development of robust and efficient speech enhancement systems.

[1]`https://sp-uhh.github.io/ears_dataset/`



(a) *whisper* (b) *regular* (c) *loud* (d) *yelling*

Figure 1: ***High Dynamic Range.*** *The EARS dataset spans the complete dynamic range of human speech, from whispering to yelling and screaming.*

2. EARS dataset

A good speech dataset is characterized by its scale, diversity, and high recording quality. However, most existing datasets fall short in one or more of these characteristics; see Table 1. Most notably, a dataset that is of high recording quality (clean 48 kHz audio), has a sufficient scale and covers the full range of human speech as opposed to only reading or neutral speech does not exist to the best of our knowledge. Yet, such a dataset is strongly required to advance research ranging from speech synthesis over voice and style conversion to speech enhancement.

We overcome these limitations with the EARS dataset, which provides high speaker and speech diversity paired with the highest recording quality.

High Recording Quality. All speech is recorded in an anechoic chamber as 32-bit audio at 48 kHz. We simultaneously record with a low-noise GRAS 40HH and a GRAS 48BL microphone, which are both mounted about 1 m in front of the participant. The first microphone has low self-noise and high sensitivity to capture subtle and nuanced speech signals, while the second has lower sensitivity to capture high-energy speech like yelling without clipping, allowing us to capture the full dynamic range of human speech, see Figure 1. We use the high-sensitivity recording for our dataset whenever possible. In the few (5% of the dataset) cases, like yelling, where the high-sensitivity microphone clips, we replace it with the lower-sensitivity microphone. To maintain the same audio characteristics between both microphones, we measure the transfer function between them using a sine-sweep and deconvolution and equalize the low-sensitivity microphone accordingly. See the project page for examples[1].

High Speaker Diversity. We recorded 107 speakers from diverse demographic backgrounds, each for close to one hour, resulting in a dataset with 100 h of clean speech. Our speakers

	hours	speakers	sample rate	anechoic	reading styles	freeform speech	emotional speech	speaker metadata
DNS (LibriVox) [10]	556	1948	48 kHz†	✗	n/a	✗	✗	✗
MSP-Conversation (v1.0) [11]	14	197	16 kHz	✗	n/a	✓	✓	✗
MSP-Podcast (v1.10) [12]	166	1458	16 kHz	✗	n/a	✓	✓	✗
LibriSpeech [4]	982	2484	16 kHz	✗	n/a	✗	✗	✗
LJSpeech [13]	24	1	22.05 kHz	✗	neutral	✗	✗	✗
TIMIT [14]	5	632	16 kHz	✗	neutral	✗	✗	✓
VCTK [5]	44	110	48 kHz	✗	neutral	✗	✗	✗
WSJ0 [15]	29	119	16 kHz	✗	neutral	✗	✗	✗
EARS (ours)	100	107	48 kHz	✓	7 styles	✓	22 emotions	✓

Table 1: ***Speech datasets.*** *In contrast to existing datasets, the EARS dataset is of higher recording quality, large, and more diverse. Reading tasks feature seven styles (regular, loud, whisper, fast, slow, high pitch, and low pitch). Additionally, the dataset features unconstrained freeform speech and speech in 22 different emotional styles. We provide transcriptions of the reading portion and metadata of the speakers (gender, age, race, first language).* †*contains files with limited bandwidth*

	# files	rooms
ACE-Challenge [16]	84	building lobby, lecture room, meeting room, office
AIR [17]	344	auditorium, corridor, lecture room, meeting room, stairway
ARNI [18]	1000	variable acoustics laboratory
BRUDEX [19]	144	variable acoustics laboratory
dEchorate [20]	648	variable acoustics laboratory
DetmoldSRIR [21]	49	concert hall, music chamber, theater
Palimpsest [22]	44	air raid shelter, dockyard, submarine

Table 2: ***RIR datasets***. *To construct EARS-Reverb, we use 2313 RIR files with different room characteristics.*

range from age 18 to 75 and span various ethnicities, including African American, Caucasian, Hispanic, and Asian. Participants are 44% male, 53% female, and 3% non-binary.

High Content Diversity. Each speaker follows a script that covers a wide variety of speech styles. The script contains a large portion of phonetically balanced sentence reading in seven different styles (regular, loud, whisper, fast, slow, low pitch, and high pitch). Additionally, it contains 18 minutes of conversational freeform speech, where participants freely reply to open-ended questions asked by an operator or talk about vacations, hobbies, or professions. To cover the wide range of emotional speech, we ask participants to read three sentences and describe an image with a specific emotional tone for each of 22 different emotions, including base emotions like ecstasy, fear, anger, or sadness, and nuanced emotions like serenity or adoration. To cover the full variety of human sounds, we additionally include short sections with non-speech sounds like laughter, yelling, or crying, vegetative sounds like coughing or yawning, interjection words, and melodic sounds. A trained operator monitors the participant during the recordings and ensures that speaking styles and prompts are followed as requested and re-record faulty segments.

3. Benchmarks

The EARS dataset enables various speech processing tasks to be evaluated in a controlled and comparable way. Here, we present benchmarks for speech enhancement and dereverberation tasks. According to the typical convention, we divide the data into training, validation, and test splits. We select participants *p001* to *p099* for training, *p100* and *p101* as validation speakers, and *p102* to *p107* as test speakers. We use all speech files except utterances containing interjection, melodic, nonverbal, or vegetative sounds. We cut longer files in the validation and training splits every 10 s to be at least 4 s long. For the test set, we provide cutting times and exclude files that are longer than 29 s. This results in 32,485 files (86.8 h) for training, 632 files (1.7 h) for validation, and 886 files (3.7 h) for the test. Data generation scripts can be found online[1].

	pub. date	# params	GMACs	proc/s [s]
Conv-TasNet [6]	May 2019	8.7 M	28	0.015
CDiffuSE [8]	May 2022	18.1 M	18,382	42.268
Demucs [7]	June 2023	83.6 M	60	0.027
SGMSE+ [9]	June 2023	64.8 M	47,984	2.575

Table 3: ***Baseline methods***. *Date of publication, number of parameters, MACs for an input of four seconds, and average processing time per one-second input length.*

3.1. EARS-WHAM

For the speech enhancement task, we construct the EARS-WHAM dataset, which mixes speech from the EARS dataset with real noise recordings from the WHAM! dataset [23] (CC BY-NC 4.0 license). We mix speech and noise files at signal-to-noise ratios (SNRs) randomly sampled in a range of $[-2.5, 17.5]$ dB, where we compute the SNR using loudness K-weighted relative to full scale (LKFS) standardized in ITU-R BS.1770 [24] to obtain a more perceptually meaningful scaling and also to remove silent regions from the SNR computation [25]. We additionally create a **blind test set** for which we only publish the noisy audio files but not the clean ground truth. It contains 743 files (2 h) from six speakers (3 male, 3 female) that are not part of the EARS dataset and noise especially recorded for this test set. We set up an evaluation server for blind evaluation on this test set, which can be found online[1].

3.2. EARS-Reverb

For the task of dereverberation, we use real recorded room impulse responses (RIRs) from multiple public datasets [16, 17, 18, 19, 20, 21, 22] (CC BY 4.0, MIT license). Table 2 shows statistics on the RIR datasets used. All RIRs are fullband, and we use a randomly selected channel for multi-channel recordings. We generate reverberant speech by convolving the clean speech with the RIR. To avoid a time delay between the reverberant and clean speech signal caused by the direct path of the RIR, we cut off the beginning of the RIR up to the index with the highest amplitude. We only use RIRs with an RT_{60} rever-

	POLQA (14 kHz)	SI-SDR [dB] (24 kHz)	PESQ (7 kHz)	ESTOI (5 kHz)	SIGMOS (24 kHz)	DNSMOS (8 kHz)	WER [%] (8 kHz)
Noisy	1.71 ± 0.56	5.98 ± 6.10	1.24 ± 0.22	0.49 ± 0.15	1.95 ± 0.39	2.74 ± 0.29	33 ± 29
Conv-TasNet [6]	2.73 ± 0.78	$\mathbf{16.93 \pm 4.36}$	2.31 ± 0.59	0.70 ± 0.14	2.69 ± 0.42	3.47 ± 0.31	20 ± 20
CDiffuSE [8]	1.81 ± 0.50	8.35 ± 3.13	1.60 ± 0.40	0.53 ± 0.15	2.08 ± 0.31	2.87 ± 0.26	32 ± 27
Demucs [7]	2.97 ± 0.75	16.92 ± 4.35	2.37 ± 0.58	0.71 ± 0.14	2.87 ± 0.43	3.66 ± 0.30	17 ± 18
SGMSE+ [9]	$\mathbf{3.40 \pm 0.73}$	16.78 ± 4.47	$\mathbf{2.50 \pm 0.62}$	$\mathbf{0.73 \pm 0.13}$	$\mathbf{3.41 \pm 0.41}$	$\mathbf{3.88 \pm 0.26}$	$\mathbf{16 \pm 18}$

Table 4: ***Results on EARS-WHAM**. Column groups are the method name, intrusive metrics, non-intrusive metrics, and WER. Below each metric is the maximum frequency taken into account for the assessment. Values indicate mean and standard deviation.*

	POLQA	SI-SDR [dB]	PESQ	ESTOI	SIGMOS	DNSMOS	WER [%]
Noisy	1.81 ± 0.60	6.48 ± 6.76	1.28 ± 0.32	0.57 ± 0.18	1.97 ± 0.44	2.79 ± 0.37	28 ± 25
Conv-TasNet [6]	2.68 ± 0.75	16.56 ± 5.80	2.41 ± 0.63	0.75 ± 0.14	2.70 ± 0.38	3.43 ± 0.35	23 ± 22
CDiffuSE [8]	1.93 ± 0.61	8.22 ± 3.97	1.64 ± 0.46	0.59 ± 0.17	2.09 ± 0.34	2.92 ± 0.29	31 ± 25
Demucs [7]	3.03 ± 0.79	$\mathbf{16.81 \pm 5.94}$	2.50 ± 0.63	0.76 ± 0.14	2.82 ± 0.43	3.62 ± 0.34	19 ± 20
SGMSE+ [9]	$\mathbf{3.35 \pm 0.82}$	16.43 ± 6.12	$\mathbf{2.59 \pm 0.70}$	$\mathbf{0.78 \pm 0.13}$	$\mathbf{3.30 \pm 0.40}$	$\mathbf{3.79 \pm 0.30}$	$\mathbf{19 \pm 19}$

Table 5: ***Results for the blind test**. Column groups are the method name, intrusive metrics, non-intrusive metrics, and WER. Values indicate mean and standard deviation.*

beration time that does not exceed 2 s. Finally, we normalize the loudness of the reverberant speech to the loudness of the clean speech using LKFS.

4. Baselines and Evaluation

4.1. Baselines

Table 3 shows all baseline methods with the date of publication, number of parameters, multiply–accumulate operations (MACs) for an input of 4 s using the ptflops package[1], and the processing time per input second. We calculate the processing time per second averaged over 20 utterances from the test set using an NVIDIA RTX A6000 graphics processing unit (GPU).

Conv-TasNet [6] is a predictive method initially proposed for speech separation that operates in the time domain. Identical to the original approach, we learn 2 ms filters, which correspond to kernels of size 120 and stride of 60 at a sampling rate of 48 kHz. We train with a batch size of 4 using one GPU.

CDiffuSE [8] is a generative speech enhancement method based on a conditional diffusion process defined in the time domain. We adapt the method for 48 kHz by using a 3072-point short-time Fourier transform (STFT), resulting in 1537 frequency bins for the conditioner. We train the large model with a batch size of 16 using two GPUs.

Demucs v4 [7] is a predictive model originally proposed for music separation. We train with batch size 8 using one GPU.

SGMSE+ [9] is a generative speech enhancement method based on a conditional diffusion process defined in the complex STFT domain. We adapt the method for 48 kHz by using 1534-point STFT with hop size 384. We use $\alpha = 0.667$ and $\beta = 0.065$ for the STFT amplitude compression and $\sigma_{\text{min}} = 0.1$, $\sigma_{\text{max}} = 1$, and $\gamma = 2$ for the stochastic differential equation. We train with a batch size of 4 using four GPUs.

4.2. Metrics

We employ *intrusive* metrics that rate the processed signal in relation to the clean reference signal and *non-intrusive* metrics, which assess the performance only using the processed signal.

Intrusive metrics include the perceptual objective listening quality analysis (POLQA) [26] for predicting speech quality, which takes values from 1 (poor) to 5 (excellent) as usual for mean opinion scores (MOS). We report the perceptual evaluation of speech quality (PESQ) [27], which is the predecessor of POLQA and is still widely used in the research community. The PESQ score lies between 1 (poor) and 4.5 (excellent). We further use extended short-time objective intelligibility (ESTOI) [28] as an intrusive measure of speech intelligibility. This metric yields values between 0 and 1, with higher values indicating better intelligibility. Moreover, we calculate the scale-invariant signal-to-distortion ratio (SI-SDR) [29] measured in dB, with higher values indicating better performance.

Non-intrusive metrics include the SIGMOS estimator [30], which is a speech quality assessment model based on a multi-dimensional listening test [31]. We report the overall quality (SIGMOS) and the reverberation assessment (MOS Reverb, only in Table 7). In addition, we use the speech quality assessment model DNSMOS [32] that is trained on human ratings obtained from listening experiments based on ITU-T P.808 [33].

To evaluate the effect of speech enhancement on automatic speech recognition (ASR), we use QuartzNet15x5Base-En from the NeMo toolkit [34] as a downstream ASR system and report the word error rate (WER). We obtain the reference transcriptions by performing ASR on the clean speech utterances.

4.3. Evaluation

We provide an empirical evaluation of the speech enhancement and dereverberation benchmarks. Listening examples for both tasks can be found online[1].

Speech Enhancement. In Table 4 and Table 5, we show speech enhancement results on the EARS-WHAM test set and the blind test set, respectively. Among the methods, the generative SGMSE+ [9] performs the best across most metrics, with particularly high scores in POLQA and SIGMOS. Demucs [7], as a representative of predictive methods, convinces with strong results, too, although falling slightly behind SGMSE+.

Listening Test. We conduct a MUSHRA-like (*Multiple Stimuli with Hidden Reference and Anchor*) listening test on EARS-WHAM with 20 participants. We randomly sample 10 distinct utterances from the test set in a gender-balanced way (5 male, 5 female). We use the clean audio as the hidden refer-

[1] https://pypi.org/project/ptflops/

	(a) POLQA				(b) SI-SDR [dB]			
	0 dB	5 dB	10 dB	15 dB	0 dB	5 dB	10 dB	15 dB
Noisy	1.2 ± 0.2	1.4 ± 0.3	1.9 ± 0.4	2.4 ± 0.4	-1.6 ± 2.4	3.5 ± 2.2	8.6 ± 2.6	13.5 ± 2.2
Conv-TasNet [6]	1.9 ± 0.5	2.5 ± 0.5	3.1 ± 0.5	3.5 ± 0.5	11.7 ± 2.6	15.2 ± 2.0	$\mathbf{19.0 \pm 2.0}$	$\mathbf{21.8 \pm 1.7}$
CDiffuSE [8]	1.3 ± 0.2	1.6 ± 0.3	2.0 ± 0.4	2.3 ± 0.5	4.5 ± 1.9	8.1 ± 1.9	10.1 ± 2.0	10.7 ± 2.1
Demucs [7]	2.1 ± 0.5	2.7 ± 0.5	3.4 ± 0.4	3.7 ± 0.3	$\mathbf{11.9 \pm 2.6}$	$\mathbf{15.3 \pm 2.3}$	18.9 ± 2.3	21.7 ± 2.1
SGMSE+ [9]	$\mathbf{2.6 \pm 0.6}$	$\mathbf{3.3 \pm 0.6}$	$\mathbf{3.8 \pm 0.3}$	$\mathbf{4.0 \pm 0.3}$	11.6 ± 2.7	15.1 ± 2.2	18.8 ± 2.2	21.8 ± 2.0

	(c) ESTOI				(d) WER [%]			
	0 dB	5 dB	10 dB	15 dB	0 dB	5 dB	10 dB	15 dB
Noisy	0.32 ± 0.08	0.44 ± 0.09	0.56 ± 0.10	0.65 ± 0.11	63 ± 25	39 ± 24	18 ± 17	12 ± 16
Conv-TasNet [6]	0.58 ± 0.14	0.67 ± 0.12	0.75 ± 0.11	0.79 ± 0.10	39 ± 22	21 ± 18	11 ± 12	8 ± 12
CDiffuSE [8]	0.37 ± 0.10	0.50 ± 0.11	0.60 ± 0.11	0.65 ± 0.11	61 ± 24	36 ± 22	18 ± 16	14 ± 18
Demucs [7]	0.60 ± 0.13	0.69 ± 0.11	0.76 ± 0.10	0.80 ± 0.10	32 ± 21	$\mathbf{17 \pm 16}$	9 ± 11	7 ± 11
SGMSE+ [9]	$\mathbf{0.63 \pm 0.13}$	$\mathbf{0.72 \pm 0.11}$	$\mathbf{0.78 \pm 0.10}$	$\mathbf{0.81 \pm 0.10}$	$\mathbf{30 \pm 21}$	$\mathbf{17 \pm 16}$	$\mathbf{8 \pm 9}$	$\mathbf{6 \pm 10}$

Table 6: ***Results per input SNR***. *Mean and standard deviation for (a) POLQA, (b) SI-SDR, (c) ESTOI, and (d) WER on EARS-WHAM.*

	POLQA	SI-SDR [dB]	PESQ	ESTOI	SIGMOS	MOS Reverb	WER [%]
Reverberant	1.75 ± 0.48	-16.17 ± 9.77	1.48 ± 0.37	0.52 ± 0.17	2.77 ± 0.43	2.99 ± 0.74	25 ± 25
SGMSE+ [9]	$\mathbf{3.61 \pm 0.63}$	$\mathbf{5.79 \pm 7.97}$	$\mathbf{3.03 \pm 0.67}$	$\mathbf{0.85 \pm 0.09}$	$\mathbf{3.49 \pm 0.43}$	$\mathbf{4.73 \pm 0.21}$	$\mathbf{9 \pm 12}$

Table 7: ***Results on EARS-Reverb***. *Column groups are the method name, intrusive metrics, non-intrusive metrics, and WER in percent. Values indicate mean and standard deviation.*

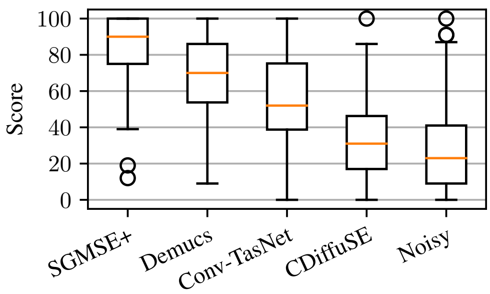


Figure 2: ***Results of the listening test***. *Subjective scores based on 20 participants visualized in a standard box plot.*

ence and the noisy audio as the hidden anchor. As stimuli, we use enhanced files of each noisy utterance from the four methods. We present participants with six audio files (four stimuli, hidden reference, hidden anchor) per utterance and ask them to "rate the overall quality considering artifacts and residual noise" of each on a scale of 0–100. The trends support the quantitative evaluation, demonstrating that SGMSE+ [9] is the preferred approach, closely followed by Demucs [7], see Figure 2.

Effect of Input SNR. Table 6 shows POLQA, SI-SDR, ESTOI, and WER scores segmented by input SNR, where 0 dB denotes the range $[-2.5, 2.5]$ dB, and each subsequent 5 dB increment representing the next range. As expected, there is a trend for better performance at higher input SNR, as well as smaller standard deviations than on the full test set.

Effect of Speaking Style and Emotion. We compare the performance of all baseline methods with respect to speaking style and selected core emotions in Table 8 and 9. We observe worse performance for whispered speech, which is expected since such voiceless speech is particularly difficult to recover after contamination with noise. Furthermore, it can be seen that all considered approaches trained on EARS-WHAM generalize well to emotional speech.

	regular	whisper	loud	slow	fast
Noisy	1.74	1.85	1.75	1.68	1.72
Conv-TasNet [6]	2.82	2.49	2.79	2.75	3.17
CDiffuSE [8]	1.78	1.68	1.93	1.72	2.02
Demucs [7]	2.95	2.82	3.10	2.86	3.27
SGMSE+ [9]	3.39	2.89	3.70	3.33	3.64

Table 8: ***POLQA for different speaking styles***. *Mean values.*

	neutral	anger	desire	pain	relief
Noisy	1.77	1.61	1.65	1.86	1.74
Conv-TasNet [6]	2.78	2.53	2.71	2.72	2.76
CDiffuSE [8]	1.65	1.94	1.76	1.94	1.78
Demucs [7]	2.93	2.91	2.82	3.02	2.91
SGMSE+ [9]	3.18	3.45	3.15	3.40	3.28

Table 9: ***POLQA for different emotions***. *Mean values.*

Dereverberation. Blind dereverberation with only a single microphone is known to be challenging, and recent results suggest that generative approaches are particularly well suited for this task [35]. In Table 7, we show dereverberation results on the EARS-Reverb test set, using the diffusion-based generative model SGMSE+ [9].

5. Conclusion

We released EARS, a dataset with high speaker and speaking style diversity spanning the full range of human speech. We hope this dataset will serve the community as a useful source to tackle new frontiers in speech processing. We additionally provided a speech enhancement and dereverberation benchmark on this new large-scale dataset and compared predictive and generative baselines to set a standard for future speech enhancement work on EARS.

6. Acknowledgements

This work has been funded by the German Research Foundation (DFG) in the transregio project Crossmodal Learning (TRR 169) and DASHH (Data Science in Hamburg – Helmholtz Graduate School for the Structure of Matter) with Grant-No. HIDSS-0002. We would like to thank J. Berger and Rohde & Schwarz SwissQual AG for their support with POLQA.

7. References

[1] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[2] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

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

[5] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[8] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7402–7406.

[9] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.

[10] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, H. Gamper, M. Golestaneh, and R. Aichner, "ICASSP 2023 deep noise suppression challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[11] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *ISCA Interspeech*, 2020, pp. 1823–1827.

[12] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2019.

[13] K. Ito and L. Johnson, "The LJ Speech Dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[14] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[15] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete - Linguistic Data Consortium," 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93s6a

[16] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[17] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *IEEE Int. Conference on Digital Signal Processing*, 2009.

[18] K. Prawda, S. J. Schlecht, and V. Välimäki, "Robust selection of clean swept-sine measurements in non-stationary noise," *The Journal of the Acoustical Society of America*, vol. 151, no. 3, pp. 2117–2126, 2022.

[19] D. Fejgin, W. Middelberg, and S. Doclo, "BRUDEX database: Binaural room impulse responses with uniformly distributed external microphones," in *Proc. ITG Conference on Speech Communication*, 2023, pp. 126–130.

[20] D. D. Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, and S. Gannot, "dEchorate: a calibrated room impulse response dataset for echo-aware signal processing," *EURASIP Journal on Audio, Speech, and Music Processing*, 2021.

[21] S. V. Amengual Gari, B. Sahin, D. Eddy, and M. Kob, "Open database of spatial room impulse responses at Detmold university of music," in *Audio Engineering Society Convention 149*, 2020.

[22] "A sonic Palimpsest: Revisiting Chatham historic dockyards." [Online]. Available: https://research.kent.ac.uk/sonic-palimpsest/impulse-responses/

[23] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending speech separation to noisy environments," in *ISCA Interspeech*, 2019, pp. 1368–1372.

[24] Recommendation ITU-R BS.1770-5, "Algorithms to measure audio programme loudness and true-peak audio level," *International Telecommunication Union (ITU)*, 2023. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1770-5-202311-I/en

[25] C. J. Steinmetz and J. D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *150th AES Convention*, 2021.

[26] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," *Int. Telecom. Union (ITU)*, 2018. [Online]. Available: https://www.itu.int/rec/T-REC-P.863-201803-I/en

[27] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001, pp. 749–752.

[28] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[29] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 626–630.

[30] N. C. Ristea, A. Saabas, R. Cutler, B. Naderi, S. Braun, and S. Branets, "ICASSP 2024 speech signal improvement challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[31] B. Naderi, R. Cutler, and N.-C. Ristea, "Multi-dimensional speech quality assessment in crowdsourcing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[32] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 6493–6497.

[33] ITU-T Rec. P.808, "Subjective evaluation of speech quality with a crowdsourcing approach," *International Telecommunication Union*, 2021. [Online]. Available: https://www.itu.int/rec/T-REC-P.808-202106-I/en

[34] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "NeMo: a toolkit for building AI applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.

[35] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.