Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Evaluating Speech Enhancement Performance Across Demographics and Languages

Jose Giraldo[1], *Alex Peiro-Lilja*[1,2], *Carme Armentano-Oller*[1], *Rodolfo Zevallos*[1], *Cristina España-Bonet* [1,3]

[1]Langtech Lab, Barcelona Supercomputing Center, Spain; [2]Centre de Llenguatge i Computació, Universitat de Barcelona, Spain; [3]DFKI GmbH, Saarland Informatics Campus, Germany

{jose.giraldo,alexandre.peiro,carme.armentano,rodolfo.zevallos,cristina.espana}@bsc.es

Abstract

Speech enhancement models have traditionally relied on VoiceBank-DEMAND for training and evaluation. However, this dataset presents significant limitations due to its limited diversity and simulated noise conditions. As an alternative, we propose and demonstrate the usefulness of evaluating the generalization capabilities of recent speech enhancement models using CommonPhone, a multilingual and crowdsourced dataset. Since CommonPhone is derived from CommonVoice, it allows to analyze enhancement performance based on demographic variables such as age and gender. Our experiments reveal significant performance variations across these variables. We also introduce a new benchmark dataset designed to challenge enhancement models with difficult and diverse speech samples, facilitating future research in universal speech enhancement.
Index Terms: speech enhancement, evaluation, multilingual

1. Introduction

The task of speech enhancement has a wide range of real-world applications. Some examples include real-time enhancement for video conferences, audio restoration of historical recordings as well as hearing aid and intelligibility improvement. For such applications, it is highly desirable to assess the generalization of the models to a worldwide population. However, the majority of speech enhancement models are benchmarked on the widely known dataset Voicebank-DEMAND (VB-DMD) [1] which has a limited diversity on demographic variables such as age and language. The training set contains only 28 speakers (14 per gender) with ages predominantly between 20-24 years, except for one 38-year-old male speaker. The test set is even more restricted, containing just two speakers aged 23-24 years. While alternative datasets such as DNS [2, 3] and CHiME [4, 5, 6] improve upon VB-DMD by offering greater speaker diversity and more varied noise conditions, they are restricted exclusively to English. Similarly, the EARS dataset [7] addresses some of VB-DMD's limitations by providing approximately 100 hours of recordings with 111 speakers covering a broader age range (18 to 75 years). Although these datasets enhance several key aspects, such as speaker age diversity, acoustic environment variety, and the amount of training data available, they remain monolingual, and the noisy examples continue to follow the simulation paradigm.

To overcome these limitations, CommonPhone (CP) [8] provides a diverse, multilingual, crowdsourced dataset that allows speech enhancement models to be evaluated beyond English. In addition to offering linguistic diversity, its rich demographic metadata enables analysis of model performance across different age groups and genders, while its phoneme-level annotations facilitate studies on content preservation. However, despite its advantages, CP has not yet been widely adopted for benchmarking speech enhancement models.

In this paper, we leverage CommonPhone to conduct a comprehensive evaluation of speech enhancement models, assessing their generalization across languages and demographic groups. By analyzing performance variations related to age and gender, as well as examining content preservation through phoneme-level annotations, we provide deeper insights into potential biases and the impact of enhancement methods on speech intelligibility and accuracy. Additionally, we extend CP by adding Catalan, the language with the most hours of recorded speech on CommonVoice to date.

Our study introduces three major contributions: (1) establishes a systematic evaluation of speech enhancement models on a demographically diverse, multilingual dataset, (2) provides detailed insights into model generalization across different speaker groups and languages, and (3) offers a new language, gender and age balanced dataset to benchmark speech enhancement in real world scenarios. These findings will be crucial for developing more robust and inclusive speech enhancement systems that perform consistently across diverse speaker populations and linguistic contexts.

1.1. Universal speech enhancement

Past works [9, 10] addressing Universal speech enhancement have focused on models that can handle a wide range of audio distortions. Following the same direction, the research community has recently proposed the URGENT challenge [11] to evaluate models in more realistic conditions with variations of sampling rates and distortions. Although it started using only English, the 2025 edition has expanded its scope to include data from 5 languages, thereby improving the assessment of model generalization under more diverse conditions. Unlike previous benchmarks such as the DNS Challenge [2, 3], the CHiME Challenge [4, 5, 6], and the Clarity Challenge [12], which have focused on specific tasks such as noise suppression, speech enhancement in reverberant environments, and intelligibility improvement for hearing aids, the new challenge takes a broader approach by adding bandwidth extension and clipping removal. However, true universality requires not only addressing various types of degradation but also ensuring robustness across speaker characteristics. While the performance of enhancement models across different audio conditions has been extensively studied, their behavior with respect to speaker diversity has received limited attention. Only recently multilingual test sets have been added [13], yet non-English languages remain underrepresented in evaluation frameworks.

10.21437/Interspeech.2025-1760

2. Experimental Setup

2.1. Dataset

We use the full CommonPhone (CP) dataset (train, dev, and test) for evaluation. CP contains 76,307 speech samples from six different languages: English (en), French (fr), Italian (it), Spanish (es), German (de), and Russian (ru). The dataset comprises recordings from 11,246 unique speakers, amounting to 116.5 hours of speech. To extend our study, we introduce a new language, Catalan, by incorporating the CommonVoice (CV) Benchmark Catalan Accents dataset.[1] This dataset consists of 16,405 audio samples from 1,531 speakers, totaling 25 hours of audio. Like CP, it is also gender balanced and sourced from CV.

By merging both datasets, we create a joint benchmarking corpus with a final duration of 141.5 hours. Given that CV captures a broad spectrum of speaker profiles and real-world acoustic conditions, it presents an ideal scenario for testing universal audio enhancement strategies. The dataset includes various types of audio degradations, such as band limitation, loudness and dynamic variations, equalization differences due to recording setups, additive background or electrical noise, and reverberation effects. Additionally, it contains common speech artifacts such as clipping, plosiveness, and sibilance, further enriching the challenges faced by speech processing models.

2.2. Models

Model selection for evaluation was based on multiple criteria. We selected the top 4 models w.r.t. PESQ[14] performance on the VB-DMD dataset, prioritizing those with open source implementations and VB-DMD pretraining. These models were chosen to ensure architectural diversity across different training paradigms. For comparison with signal processing approaches, we included a spectral gate baseline implemented in noisereduce [15]. See Table 1 for descriptions of all selected models.

Table 1: *Models selected for benchmarking.*

Model	Params.	Type of Training	Input
SEMamba[16]	2.25M	Adversarial	Mag+phase
Mp-Senet[17]	2.05M	Adversarial	Mag+phase
Openuniverse++[18]	84.24M	Adversarial+Diffusion	Waveform+Mel
SGMSE+[19]	65.59M	Diffusion	Complex

2.3. Metrics

2.3.1. Speech quality metrics

Following evaluation pipelines of previous works, PESQ [14], STOI [20] and SI-SDR [21] are chosen. Due to the lack of clean reference audio, we estimate these metrics with the SQUIM model [22]. PESQ is useful for the evaluation of generative speech enhancement in low-SNR [23]. Considering the extensive scale of the evaluation dataset, conducting a comprehensive subjective listening assessment is infeasible. Instead, we include UTMOS [24] , SCOREQ [25] and NISQA [26] which have been shown to correlate well with MOS ratings.[2]

[1] doi:10.57967/hf/5679

[2] Opening Remarks Urgent 2024 Challenge `https://neurips.cc/virtual/2024/102916`

2.3.2. Content metrics

To analyze the capacity of enhancement models to maintain speech information, we computed Word Information Loss (WIL) and Word Error Rate (WER) to evaluate the predicted transcriptions from enhanced samples. WIL improves upon WER by weighting errors based on their impact on meaning, aligning better with human perception. Equation 1 describes WIL, where N and P represent the total number of target and predicted words, respectively, while C denotes the number of correct words.

$$WIL = 1 - \frac{C}{N} * \frac{C}{P} \tag{1}$$

The predicted transcriptions were generated using Conformer-CTC Large speech-to-text models from NVIDIA NeMo's platform,[3] which provides pre-trained models for all the languages explored in this work and does not rely on a language model for corrections. Moreover, we were able to compute Phoneme Error Rate (PER) to evaluate the samples at a more fine-grained level. The original Wav2Vec2 model fine-tuned on the six CP languages for phoneme recognition from [8] was performed on enhanced samples. For Catalan, we applied Facebook's fine-tuned multilingual Wav2Vec2[4] to predict phoneme sequences, which were later compared with those obtained from the eSpeak[5] rule-based phonemizer.

2.3.3. Linguistic distance metrics

To compare the performance of speech enhancement models in different languages with their proximity to English, we quantified the distance between languages using the URIEL+ library [27]. This tool represents languages through various features, and allows the calculation of the distance between languages considering multiple linguistic factors.

Table 2: *Language distances to English, lower is nearer.*

Distance	de	es	fr	it	ru	ca
Phonological	0.37	0.28	0.46	(no data)	0.28	0.20
Phonemic Inv.	0.44	0.55	0.48	0.51	0.56	0.46
Phylogenetic	0.64	0.94	0.94	0.93	0.89	0.93

For this study, we specifically considered Phonological distance, Phonemic Inventory distance and Phylogenetic distance. Phonological Distance refers to the characteristics in the sound systems of languages, covering both segmental and suprasegmental features. In contrast, Phonemic Inventory Distance is based on the set of phonemes for each language. Phylogenetic distance refers to the shared membership in language families, based on the world language family tree in Glottolog [28]. The distances considering these characteristics between English and the analyzed languages are reported in Table 2.

3. Evaluation

3.1. Demographic analysis

3.1.1. Age and gender

For the analysis, we discard speakers older than 79 years old as there are only 10 English speakers in that age band. The rest

[3] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_(ca, de, en, es, fr, it, ru)_conformer_ctc_large

[4] https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft

[5] https://github.com/espeak-ng/espeak-ng

Table 3: *Comparison of models in terms of information loss (L), PER, and WIL by language, with all scores in percentage.*

	en			ca			de			es			fr			it			ru		
	L	**PER**	**WIL**	**L**	**PER**	**WIL**	**L**	**PER**	**WIL**	**L**	**PER**	**WIL**	**L**	**PER**	**WIL**	**L**	**PER**	**WIL**	**L**	**PER**	**WIL**
MPS	24.5	38	36.6	20.0	40	32.3	29.2	32	34.9	16.5	32	33.6	36.7	32	40.1	17.7	35	37.4	31.5	21	30.7
OpU	40.1	54	46.0	33.4	50	41.2	46.7	46	43.8	37.6	46	42.9	52.0	46	48.0	39.9	48	46.6	47.2	37	42.5
SMb	**23.8**	**37**	35.2	**18.8**	**39**	**31.3**	**28.1**	**31**	**33.5**	**14.5**	**27**	29.6	**36.0**	**30**	**38.6**	**15.7**	**30**	32.9	**30.6**	**20**	**30.3**
SGM	29.6	42	39.1	27.0	43	36.2	35.3	35	37.1	24.6	35	34.9	41.4	35	41.4	28.0	38	39.2	37.7	28	35.6
SpGt	26.4	42	**35.0**	22.7	43	32.9	35.9	34	33.8	19.1	27	**27.0**	38.9	34	38.8	20.5	31	**30.6**	37.5	26	32.6

of the languages have speakers in all of the age bins, except for Russian that only has speakers up to 50 years old. Figure 1 shows a decrease in performance as age increases for both NISQA and SCOREQ metrics. In the case of UTMOS, while this decrease is less pronounced for the best model (SGMSE+), the remaining models maintain the trend of decreasing performance. For UTMOS and SCOREQ, although the Baseline (spectral gate) method shows slight fluctuations, it does not exhibit a decreasing trend, which aligns with expectations given its non-data dependency. Openuniverse++ is also the model with the highest decrease in performance w.r.t age in SCOREQ.

Statistical analysis using Kruskal-Wallis tests, conducted separately for each model and language, confirmed significant ($\chi^2(6) = 22.46$, $p < .001$) age-related differences across all three metrics (NISQA, SCOREQ, and UTMOS). Post-hoc comparisons using Conover revealed that there are several age groups for SGMSE+ and spectral gate where no significant difference was found in UTMOS and NISQA, suggesting that the performance of those models is less affected by the age variable. Finally, the best performing age group is 20-40, which coincides with the age range of the train dataset (VB-DMD).

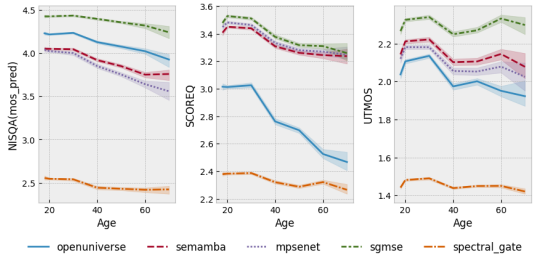

Figure 1: *Comparison of speech enhancement systems per age groups averaged over all languages. 95% CI on shaded region.*

We next examine whether speech enhancement performance differs by speaker gender, in Figure 2 the value of the NISQA, SCOREQ and UTMOS is always higher for males on SEMamba, MP-Senet and SGMSE+. The greatest differences are observed in the UTMOS metric, even for spectral gate. A Mann-Whitney U test per each model and age band was performed to evaluate whether NISQA, UTMOS and SCOREQ differed by gender. The results indicated that males had significantly ($p < 0.001$) higher values than females, except for Openuniverse++ on NISQA and SCOREQ, as well as Spectral gate on SCOREQ.

3.1.2. Language

Table 4 presents the performance metrics across different models and languages. Although STOI was also computed we the results on the table because the values are the same (0.96) for all rows. As expected, English, that is the same language of the

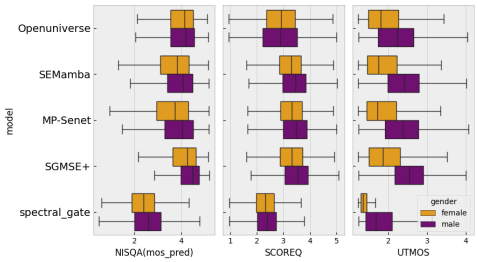

Figure 2: *Comparison of speech enhancement systems per gender groups.*

training dataset, has the highest scores in the non-reference metrics (SCOREQ and UTMOS). However, for PESQ and SI-SDR the best performing language alternates between Catalan, Italian, and Russian but the differences are less pronounced compared with non-reference metrics. Finally, SGMSE+ ranks first in all metrics.

Table 4: *Language performance comparison.*

Mod.	**Lan**	**SCOREQ**	**UTMOS**	**PESQ**	**SI-SDR**
Mp-Senet	ca	3.17 ± 0.01	1.93 ± 0.01	**2.97** ± 0.01	19.86 ± 0.11
	de	3.44 ± 0.01	2.29 ± 0.01	2.96 ± 0.01	20.06 ± 0.11
	en	**3.68** ± 0.01	**2.63** ± 0.01	2.96 ± 0.01	19.97 ± 0.11
	es	3.22 ± 0.01	1.95 ± 0.01	2.92 ± 0.01	19.09 ± 0.11
	fr	3.21 ± 0.01	2.11 ± 0.01	2.87 ± 0.01	20.05 ± 0.11
	it	3.26 ± 0.01	2.06 ± 0.01	2.95 ± 0.01	19.77 ± 0.10
	ru	3.42 ± 0.01	2.22 ± 0.01	2.95 ± 0.01	**20.33** ± 0.11
Openuniverse++	ca	2.78 ± 0.01	1.95 ± 0.01	2.94 ± 0.01	18.13 ± 0.10
	de	3.01 ± 0.01	2.22 ± 0.01	2.98 ± 0.01	18.47 ± 0.10
	en	**3.26** ± 0.01	**2.53** ± 0.01	2.94 ± 0.01	18.45 ± 0.10
	es	2.74 ± 0.01	1.90 ± 0.01	2.88 ± 0.01	17.53 ± 0.10
	fr	2.78 ± 0.01	2.08 ± 0.01	2.89 ± 0.01	**18.66** ± 0.10
	it	2.74 ± 0.01	1.99 ± 0.01	**2.98** ± 0.01	18.29 ± 0.10
	ru	2.94 ± 0.01	2.14 ± 0.01	2.97 ± 0.01	18.60 ± 0.10
SEMamba	ca	3.18 ± 0.01	1.95 ± 0.01	2.91 ± 0.01	19.66 ± 0.11
	de	3.40 ± 0.01	2.32 ± 0.01	2.94 ± 0.01	20.14 ± 0.11
	en	**3.64** ± 0.01	**2.65** ± 0.01	2.94 ± 0.01	20.14 ± 0.11
	es	3.19 ± 0.01	1.98 ± 0.01	2.92 ± 0.01	19.40 ± 0.11
	fr	3.17 ± 0.01	2.14 ± 0.01	2.85 ± 0.01	20.24 ± 0.10
	it	3.24 ± 0.01	2.10 ± 0.01	**2.96** ± 0.01	20.16 ± 0.09
	ru	3.40 ± 0.01	2.26 ± 0.01	2.94 ± 0.01	**20.44** ± 0.11
SGMSE+	ca	3.21 ± 0.01	2.08 ± 0.01	**3.16** ± 0.01	**20.99** ± 0.09
	de	3.46 ± 0.01	2.41 ± 0.01	3.05 ± 0.01	20.51 ± 0.11
	en	**3.73** ± 0.01	**2.75** ± 0.01	3.06 ± 0.01	20.60 ± 0.11
	es	3.26 ± 0.01	2.08 ± 0.01	3.02 ± 0.01	19.88 ± 0.11
	fr	3.24 ± 0.01	2.24 ± 0.01	2.98 ± 0.01	20.87 ± 0.10
	it	3.35 ± 0.01	2.23 ± 0.01	3.11 ± 0.01	20.93 ± 0.09
	ru	3.44 ± 0.01	2.34 ± 0.01	3.04 ± 0.01	20.89 ± 0.10

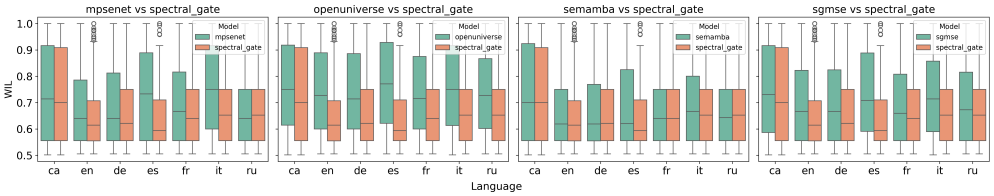

Figure 3: *Model comparison vs. spectral gating by language, for samples with WIL greater than 50%.*

3.2. Content and linguistic analysis

3.2.1. Linguistic distance

We compare the model's performance using different metrics in various languages, considering their distance from English. Regarding the distance in phonological inventory, we observe a weak negative (-0.21) correlation when considering the SCOREQ and UTMOS metrics, excluding the fact that Catalan yields anomalously low results in some of the models used and Russian anomalously high ones. This can be explained by the fact that the Catalan dataset has 30% speakers over 50 years old, while Russian does not have any (as mentioned in Section 3.1.1). The relationship between the model's performance and phylogenetic distance appears clearer with the aggregated values in Figure 4, but still yields a weak negative correlation (-0.17), especially considering that the distance differences among the Romance languages are minimal. Finally, contrary to our expectations, we do not observe a clear relationship between phonological distance and the model's performance. Nevertheless, we believe that a sample of seven languages is not significant enough to draw definitive conclusions.

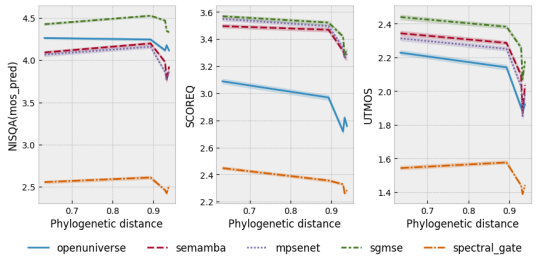

Figure 4: *Relation between speech quality metrics and phylogenetic distance to English.*

3.2.2. Information loss

Some samples showed partial or total loss of speech after being processed by the enhancement models. To analyze this phenomenon in more detail, we collected samples with a WER higher than the values reported in NeMo model evaluations for the CV test data for each language.: 9.4% (en), 4.27% (ca), 6.68% (de), 6.9% (es), 9.63% (fr), 7.2% (it), 4.3% (ru). The amount of enhanced samples that surpass these scores were considered data with information loss (L). WIL and PER were computed for these subsets, per language and model. The overall percentage results are shown in Table 3. We clearly observe that SEMamba (SMb) is the model that loses the least information across all languages. On the other hand, Openuniverse++ (OpU) performs the worst in all aspects for all languages. SGMSE+ (SGM) appears to lose more information than spectral gating (SpGt) signal processing, despite achieving the best scores in terms of audio quality.

Figure 3 presents the distributions of the worst samples per language and model (those with a WIL higher than 50%). The four enhancement models are compared with spectral gating. SGM and Mp-Senet (MPS) show varying sensitivity depending on the language, while SEMamba exhibits a lower median among the worst cases. Surprisingly, Catalan displays the largest quartiles across all models, despite being one of the languages least affected by information loss.

3.3. CommonPhone-SE

We introduce a benchmark subset of 5242 challenging speech samples to encourage more robust model development. The sampling rationale was to select audios that remain difficult for state of the art models, both in terms of speech quality metrics and content preservation, hence, we selected the worst 40 examples w.r.t. to UTMOS, SCOREQ and WIL per each language, age band and gender. Finally, the duplicates were dropped to arrive at a final evaluation dataset of 8.24 hours. The dataset is released in Huggingface [6] to facilitate access to a wide community with a webpage to share audio examples [7].

4. Discussion and future work

We found that the model ranking on VB-DMD does not directly transfer to a more diverse and realistic dataset, highlighting the limitations of using a single, simplified benchmark. SGMSE+ demonstrates superior performance in signal quality metrics and shows remarkable stability across different speaker ages. A critical finding is the potential overfitting to PESQ scores, as observed in MP-Senet and SEMamba which explicitly optimize for this metric. This aligns with recent work [29] questioning the reliability of PESQ optimization. A key takeaway is that neural models often compromise intelligibility for perceptual quality and models achieving excellent quality scores often perform poorly in content retention, as also observed in [30]. Surprisingly, for languages like Spanish and Catalan, which score lower on quality metrics, models demonstrate better content preservation. The traditional signal processing approach of spectral gating, while uncompetitive in quality metrics, outperforms neural models in information preservation in several languages. Based in the findings about spectral gate for content preservation, we believe in the potential of hybrid approaches for enhancement strategies which we will explore in the future.

Significant statistical differences across age, gender and language were found in the evaluation, supporting the need for multilingual and diverse datasets in speech enhancement evaluation and training. The released dataset will serve to the research community as tool to achieve Universal speech enhancements for a diverse population.

[6] https://huggingface.co/datasets/BSC-LT/CommonPhone-SE/
[7] https://github.com/langtech-bsc/commonphone-se

5. Acknowledgements

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project ILENIA with reference 2022/TL22/00215337 and by the Government of Catalonia through the Aina project.

6. References

[1] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, pp. 146–152.

[2] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Interspeech 2020*. ISCA, Oct. 2020.

[3] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh *et al.*, "Icassp 2023 deep noise suppression challenge," *IEEE Open Journal of Signal Processing*, 2024.

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

[5] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth'chime'speech separation and recognition challenge: dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[6] S. Leglaive, L. Borne, E. Tzinis, M. Sadeghi, M. Fraticelli, S. Wisdom, M. Pariente, D. Pressnitzer, and J. R. Hershey, "The chime-7 udase task: Unsupervised domain adaptation for conversational speech enhancement," *arXiv preprint arXiv:2307.03533*, 2023.

[7] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "Ears: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," 2024.

[8] P. Klumpp, T. Arias, P. A. Pérez-Toro, E. Noeth, and J. Orozco-Arroyave, "Common phone: A multilingual dataset for robust acoustic modelling," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 763–768. [Online]. Available: https://aclanthology.org/2022.lrec-1.81/

[9] H. Liu, X. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: A unified framework for high-fidelity speech restoration," in *Interspeech 2022*. ISCA, Sep. 2022. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2022-11026

[10] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," 2022.

[11] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt, and Y. Qian, "Urgent challenge: Universality, robustness, and generalizability for speech enhancement," in *Interspeech 2024*. ISCA, Sep. 2024, p. 4868–4872. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2024-1239

[12] M. A. Akeroyd, W. Bailey, J. Barker, T. J. Cox, J. F. Culling, S. Graetzer, G. Naylor, Z. Podwińska, and Z. Tu, "The 2nd clarity enhancement challenge for hearing aid speech intelligibility enhancement: Overview and outcomes," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] R. Cutler, A. Saabas, B. Naderi, N.-C. Ristea, S. Braun, and S. Branets, "Icassp 2023 speech signal improvement challenge," *IEEE Open Journal of Signal Processing*, vol. 5, p. 662–674, 2024. [Online]. Available: http://dx.doi.org/10.1109/OJSP.2024.3376293

[14] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.

[15] T. Sainburg, "timsainb/noisereduce: v1.0," Jun. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3243139

[16] R. Chao, W.-H. Cheng, M. La Quatra, S. M. Siniscalchi, C.-H. H. Yang, S.-W. Fu, and Y. Tsao, "An investigation of incorporating mamba for speech enhancement," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 302–308.

[17] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "Mp-senet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *INTERSPEECH 2023*. ISCA, Aug. 2023. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2023-1441

[18] R. Scheibler, Y. Fujita, Y. Shirahata, and T. Komatsu, "Universal score-based speech enhancement with high content preservation," 2024. [Online]. Available: https://arxiv.org/abs/2406.12194

[19] 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.

[20] 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.

[21] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr - half-baked or well done?" 2018. [Online]. Available: https://arxiv.org/abs/1811.02508

[22] A. Kumar, K. Tan, Z. Ni, P. Manocha, X. Zhang, E. Henderson, and B. Xu, "Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio," 2023. [Online]. Available: https://arxiv.org/abs/2304.01448

[23] J. Pirklbauer, M. Sach, K. Fluyt, W. Tirry, W. Wardah, S. Moeller, and T. Fingscheidt, "Evaluation metrics for generative speech enhancement methods: Issues and perspectives," in *Speech Communication; 15th ITG Conference*, 2023, pp. 265–269.

[24] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," 2022. [Online]. Available: https://arxiv.org/abs/2204.02152

[25] A. Ragano, J. Skoglund, and A. Hines, "Scoreq: Speech quality assessment with contrastive regression," 2025. [Online]. Available: https://arxiv.org/abs/2410.06675

[26] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Interspeech 2021*. ISCA, 2021. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2021-299

[27] A. Khan, M. Shipton, D. Anugraha, K. Duan, P. H. Hoang, E. Khiu, A. S. Doğruöz, and E.-S. A. Lee, "Uriel+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base," 2024. [Online]. Available: https://arxiv.org/abs/2409.18472

[28] H. Hammarström, "Ethnologue 16/17/18th editions: A comprehensive review," *Language*, vol. 91, no. 3, pp. 723–737, 2015, plus 188pp online appendix.

[29] D. de Oliveira, S. Welker, J. Richter, and T. Gerkmann, "The pesqetarian: On the relevance of goodhart's law for speech enhancement," 2024.

[30] I. López-Espejo, A. Edraki, W.-Y. Chan, Z.-H. Tan, and J. Jensen, "On the deficiency of intelligibility metrics as proxies for subjective intelligibility," *Speech Commun.*, vol. 150, no. C, p. 9–22, May 2023. [Online]. Available: https://doi.org/10.1016/j.specom.2023.04.001