

SP-MCQA: EVALUATING INTELLIGIBILITY OF TTS BEYOND THE WORD LEVEL

Hitomi Jin Ling Tee Chaoren Wang Zijie Zhang Zhizheng Wu

The Chinese University of Hong Kong, Shenzhen

ABSTRACT

The evaluation of intelligibility for TTS has reached a bottleneck, as existing assessments heavily rely on word-by-word accuracy metrics such as WER, which fail to capture the complexity of real-world speech or reflect human comprehension needs. To address this, we propose SP-MCQA (Spoken-Passage Multiple-Choice Question Answering), a novel subjective approach evaluating the accuracy of key information in synthesized speech, and release SP-MCQA-Eval, an 8.76-hour news-style benchmark dataset for SP-MCQA evaluation. Our experiments reveal that low WER does not necessarily guarantee high key-information accuracy, exposing a gap between traditional metrics and practical intelligibility. SP-MCQA shows that even state-of-the-art (SOTA) models still lack robust text normalization and phonetic accuracy. This work underscores the urgent need for high-level, more life-like evaluation criteria now that many systems already excel at WER yet may fall short on real-world intelligibility.

Index Terms— TTS evaluation, subjective metric, key information accuracy, benchmark dataset

1. INTRODUCTION

Text-to-Speech (TTS) [1, 2, 3, 4, 5, 6, 7, 8] systems have achieved remarkable progress in producing highly intelligible speech. However, evaluation methods have not kept pace with these advances. Existing intelligibility-related metrics, such as Word Error Rate (WER) or subjective intelligibility Mean Opinion Score (MOS), predominantly focus on low-level accuracy which overlook if the key information is accurately conveyed. This gap is critical because, in real-world scenarios, listeners care more about understanding essential information than perfect word-by-word reproduction.

Meanwhile, many existing TTS evaluation test sets [9, 10, 11] are relatively simple and standardized, lacking challenging cases that reflect real-world speech variability. Although Seed-TTS [2] introduces challenging patterns such as word repetitions and tongue twisters, it still fails to adequately assess model performance on irregular text — particularly content involving digits and proper nouns, such as locations, names, numbers, and events — which frequently constitute key information in informative, context-rich speech.

To address these shortcomings, we introduce SP-MCQA

(Spoken-Passage Multiple-Choice Question Answering), a framework for evaluating TTS systems on key-information accuracy beyond the word level, comprising a novel subjective evaluation metric (SP-MCQA ACC) and test set (SP-MCQA-Eval). In line with the need for SP-MCQA evaluation, SP-MCQA-Eval is constructed as a news-style test set for speech synthesis that is both acoustically natural and semantically rich in contextual and critical information. Unlike traditional transcript-based metrics, SP-MCQA does not measure word-by-word accuracy and instead evaluates semantic and structural fidelity of key information under realistic listening conditions. It acts as a complementary framework to WER for evaluating and comparing models that already achieve high intelligibility at the word level.

The contributions of our work are threefold:

- We propose SP-MCQA, a novel subjective evaluation approach to measure the key information accuracy of a synthesized speech.
- We create SP-MCQA-Eval, a new open-source news-style benchmark dataset that contains uncommon text, involving proper nouns and digits, designed for SP-MCQA.
- We conduct an in-depth and comprehensive evaluation of how state-of-the-art (SOTA) TTS systems perform on this benchmark.

2. RELATED WORK

Spoken Multiple-Choice Question Answering (SMCQA), a form of Spoken Question Answering (SQA), referred to tasks related to machine text comprehension in which passages, questions, and multiple choices are presented entirely in speech [12, 13]. In contrast, we study a hybrid setting similar to the text-based question answering on listening comprehension test [14], where only the passage is spoken while the questions and answer choices are in text. We term this new variant SP-MCQA (Spoken-Passage Multiple-Choice Question Answering) to clearly distinguish it from SMCQA.

There are several datasets for listening comprehension research, such as TOEFL-QA¹ and TED-Q [15]. Although these corpora resemble our newly created test set in form, they rarely focus on key information involving digits or numbers,

¹<https://github.com/iamyuanchung/TOEFL-QA>

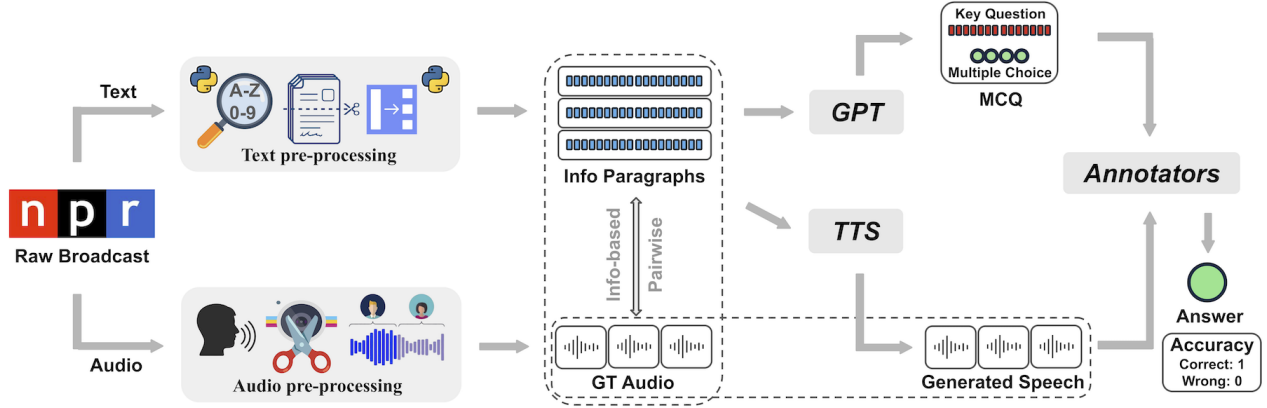


Fig. 1. An overview of SP-MCQA. The process involves two main stages: (1) The creation of SP-MCQA-Eval benchmark dataset, and (2) the pipeline for SP-MCQA evaluation.

making them less suitable than SP-MCQA-Eval for assessing the intelligibility of synthetic speech on key information.

3. SP-MCQA OVERVIEW

3.1. Benchmark Dataset

The SP-MCQA-Eval dataset contains proper nouns and digits (dates, names, times, locations, events, etc.). Table 1 summarizes the SP-MCQA-Eval statistics, comprising 5,805 utterances with a total duration of 8.76 hours of speech.

Table 1. Statistics of the SP-MCQA-Eval test set.

# Hrs	# Spk	# Para	# Utts	# Ques
8.76	483	550	5,805	2,688

Audio and manually annotated text are sourced from National Public Radio (NPR) news, providing conversational speech with rich contextual information that facilitates the extraction of key details commonly encountered in daily life.

Following Emilia’s pre-processing pipeline [16], background music is removed using Ultimate Vocal Remover (UVR)², and WhisperX [17] is used for ASR and timestamp extraction. Because uppercase letters and numbers often indicate key information, we apply Regular Expressions (RegEx) to filter the raw text, retaining paragraphs that contain: (1) at least one number with a minimum of three digits; and (2) at least two uppercase letters (excluding those at sentence beginnings). From this filtered pool, we randomly select 550 “information paragraphs,” each constrained to 65–260 words (≈ 30 s – 2 min of speech). Then, Pydub is used to segment the clean audio according to these paragraphs, producing the ground-truth for subsequent SP-MCQA testing. We further apply pyannote speaker-diarization-3.1 [18] to extract speaker timestamps and refine the segmentation to separate different

speakers. Since most TTS systems are not trained on very long utterances, we use the Natural Language Toolkit (NLTK) [19] to split each paragraph into natural sentences and segment the audio accordingly. These aligned sentence–audio pairs form the ground-truth for later objective evaluation.

We employ GPT-4o-mini [20] to automatically generate multiple-choice questions (MCQs) that examine the key information in each paragraph. Each evaluation task consists of a paragraph and its two to ten associated questions. Each MCQ contains four options: one correct answer, one “Other,” and two distractors representing different error types, as illustrated in Table 2. All questions are manually inspected, and problematic items caused by GPT hallucinations are removed.

3.2. Evaluation Pipeline

For SP-MCQA evaluation, annotators listen to speech and answer multiple-choice textual questions based on its content, with accuracy scored as 1 for a correct answer and 0 otherwise. We recruit 40 annotators who are either native English speakers or non-native with an IELTS listening score 8.0³ and above. All annotators receive clear instructions: “Select the answer that directly matches the information explicitly stated in the audio; do not infer beyond what is clearly stated.”

Each evaluation task is randomly assigned to two annotators. If their answers differ, a third annotator is added; if all three differ, a fourth is introduced. No further annotators are involved if disagreement persists among four. Golden test questions are randomly inserted into 10% of tasks to assess general knowledge. Only annotators achieving 100% accuracy on golden tests are retained; results from those below this threshold are discarded, and their tasks are reassigned to

²<https://github.com/Anjok07/ultimatevocalremovergui>

³<https://takeielts.britishcouncil.org/teach-ielts/test-information/ielts-scores-explained> Band 8.0 indicates a very good user with a fully operational command of the language, making only occasional unsystematic inaccuracies or inappropriacies, and able to handle complex, detailed argumentation with only minor misunderstandings in unfamiliar situations.

Table 2. Error type descriptions of the generated multiple-choice questions.

Error Type	Description	Example
Phonetic Error	Similar sounding to the correct answer.	“Eighteen” vs. “Eighty”
Semantic Error	Logically reasonable but factually incorrect.	“Wednesday” vs. “Thursday”
Syntax Error	Structural mistakes in phrasing.	“Ph.D. Emily Clark” vs. “Emily Clark, Ph.D.”
Grammar Error	Grammatical inconsistencies or subtle inaccuracies in expression.	“Wet cloths” vs. “Cloths wet”
Other	Always appear as an option.	“None of the above”

other qualified annotators. The final SP-MCQA ACC is computed as the average accuracy across qualified annotators. We also collect qualitative feedback for future analysis.

4. EXPERIMENTS

4.1. Models

We select FishSpeech V1.4 [5], MaskGCT [8], F5-TTS [6], and CosyVoice 2 [7] for our SP-MCQA evaluation due to their strong performance in regular intelligibility assessment. FishSpeech V1.4 employs a dual autoregressive (AR) architecture and leverages LLMs for rich linguistic feature extraction, resulting in clear pronunciation and highly intelligible output. MaskGCT, a fully non-autoregressive (NAR) model with a mask-and-predict paradigm, enables fast parallel synthesis while maintaining high word-level accuracy and strong speaker similarity across unseen speakers. F5-TTS, also fully NAR with flow-matching-based diffusion, achieves low WER and expressive speech synthesis, providing robust intelligibility and controllable speaking rates. CosyVoice 2 combines AR fidelity with NAR speed, separately modeling semantic and acoustic features to enhance prosody, rhythm, and intonation, supporting naturalness and speaker identity preservation in both streaming and non-streaming scenarios. Collectively, these models represent diverse architectures and design choices, making them ideal for assessing the accuracy, speaker consistency, and audio quality of synthesized speech in key-information-sensitive tasks under news-style content.

4.2. Evaluation Metrics

We also conduct objective evaluations on each model across three key aspects: intelligibility, coherence, and audio quality. Intelligibility is evaluated using Word Error Rate (WER), computed with the jewel package and transcribed by Whisper-large-v3⁴ ASR, with a fixed “prompt” parameter to ensure consistent transcription style. Coherence is assessed via speaker similarity (S-SIM), which computes the cosine similarity between WavLM-TDNN [21] speaker embeddings of the synthesized speech and the reference prompt. Audio quality is evaluated using Deep Noise Suppression Mean Opinion Score (DNSMOS) [22], derived from P.835 human ratings, providing an overall audio quality score on a 1-5 scale. All

metrics are computed at a 16 kHz sampling rate. While WER is our primary intelligibility metric for comparison, all three objective metrics together provide reference values for the overall quality of the SP-MCQA-Eval test set and the synthesized speech produced by TTS models.

4.3. Experimental Setup

The SP-MCQA-Eval test set contains 483 unique speakers. During inference, we select one prompt (utterance + transcript) from each speaker, while the transcripts of all utterances—including the selected prompts—serve as target texts, yielding 5,805 prompt–target pairs in total. These pairs preserve speaker identity, enabling direct evaluation against ground-truth data. TTS inferences are conducted on 8 NVIDIA GeForce RTX 4090 GPUs using the official code from each model’s GitHub repository. For MaskGCT⁵, we modify the G2P module to correctly classify numerical inputs as English (en) instead of “other language.” For CosyVoice 2⁶, we remove the duration constraint in frontend.py, allowing the model to process speech tokens for audio longer than 30 seconds. No modifications are made to the inference code for FishSpeech V1.4⁷ and F5-TTS⁸. All evaluations are also conducted under the same experimental setup as inference.

4.4. Results & Analysis

Table 3. Evaluation results of ground-truth and four TTS models on SP-MCQA-Eval test set.

System	SP-MCQA ACC (%) ↑	WER (%) ↓	S-SIM ↑	DNSMOS P.835 OVRL ↑
Ground-Truth	92.045	8.067	0.710	2.955
F5-TTS	87.139	11.267	0.654	3.202
MaskGCT	89.260	7.351	0.710	3.081
CosyVoice 2	90.399	9.044	0.523	3.334
FishSpeech	81.194	5.739	0.522	3.242

The evaluation results of ground-truth and four TTS models under SP-MCQA-Eval are shown in Table 3. The relatively moderate performance of the ground-truth mainly stem from imprecise in timestamp extraction during pre-processing; nevertheless, these values still provide a mean-

⁴<https://huggingface.com/openai/whisper-large-v3>

⁵<https://github.com/open-mmlab/Amphion/tree/main/models/tts/maskgct>

⁶<https://github.com/FunAudioLLM/CosyVoice>

⁷<https://github.com/fishaudio/fish-speech>

⁸<https://github.com/SWivid/F5-TTS>

Table 4. Analysis of error types on SP-MCQA evaluation for each system.

System	Evaluation		Error Types			
	Total Ques	Wrong Ques	Phonetic	Semantic	Structure (Syntax + Grammar)	Other
Ground-Truth	6914	550	246 (3.558%)	80 (1.157%)	49 + 61 (1.591%)	114 (1.649%)
F5-TTS	7472	961	306 (4.095%)	114 (1.526%)	79 + 93 (2.302%)	369 (4.938%)
MaskGCT	7477	803	267 (3.571%)	104 (1.391%)	74 + 93 (2.234%)	265 (3.544%)
CosyVoice 2	7218	693	233 (3.228%)	70 (0.970%)	64 + 72 (1.884%)	254 (3.519%)
FishSpeech	7519	1414	271 (3.604%)	104 (1.383%)	66 + 77 (1.902%)	896 (11.916%)

Table 5. Main comments for selecting “Other” in SP-MCQA evaluation (lightly edited for clarity).

Task ID	System	Comment	Related Issue
2210	CosyVoice 2	A “-nine” sound occurs after every sentence. (Cannot recognize).	Noise
1938	CosyVoice 2	Each sentence ends with an “-edge”/“-ged” sound, causing confusion. (Cannot recognize).	Noise
951	MaskGCT	The audio is noisy and difficult to identify. (Cannot recognize).	Noise
975	MaskGCT	In the first 55 seconds, pronunciation was unclear, resembling reversed speech or persistent difficulty articulating the governor’s name throughout the segment. (Cannot recognize).	Noise
543	MaskGCT	Pronunciation resembles “Alala” rather than “Alabama.” (No options).	Proper Noun
1156	F5-TTS	Too fast with unclear pronunciation. (Cannot recognize).	Speed
380	F5-TTS	Speech rate is approximately 1.75×. (Cannot recognize).	Speed
544	F5-TTS	Pronunciation alternates between “... Ala” and “... Alabama.” (Inconsistency).	Proper Noun
689	FishSpeech	The number is 2, not 2,000. (No options).	Number
541	FishSpeech	Only “Talladega” is audible, not “Talladega Ala” or “Talladega Alabama.” (No options).	Proper Noun
185	FishSpeech	No data was mentioned. (No options).	Number

ingful reference for evaluation. While FishSpeech ranks highest in WER, it performs worst in SP-MCQA ACC, whereas CosyVoice 2, despite its lower rank in WER, ranks highest in SP-MCQA ACC. This reveals a critical limitation of WER: models with low WER may still fail to convey key information accurately.

Table 4 presents the analysis of the error types, showing that phonetic errors are the most prevalent across all models, followed by structural (syntax and grammar) and semantic errors. The latter two might be caused by model’s “hallucinations,” such as generating inconsistent content with the input, or omitting content. We observe that NAR models (F5-TTS and MaskGCT) exhibit a higher proportion of such errors compared to AR models (CosyVoice2 and FishSpeech). Nonetheless, regardless of architecture, phonetic accuracy remains the primary challenge in our SP-MCQA tasks for key information, likely due to the scarcity of irregular or uncommon patterns in training data. These findings highlight the importance of addressing phonetic errors, especially in rare or atypical utterances, for developing human-like TTS systems.

We further analyze annotators’ comments for selecting “Other” (Table 5). CosyVoice 2 occasionally adds noise at sentence endings, raising WER and slightly lowering SP-MCQA ACC. However, it is the only model that correctly pronounces all tested abbreviations (e.g., “Ala.”→“Alabama”), achieving the highest SP-MCQA ACC. MaskGCT generates odd sounds and struggles with proper nouns (e.g., “Ala.”→“Alala”), performing slightly worse on the SP-MCQA task. F5-TTS, in addition to incorrectly handling uncommon

text (e.g., “Ala.”→“Ala”), exhibits overly fast speech, hindering the recognition of key information. FishSpeech suffers from mid-sentence word drops and cut-offs due to normalization issues (e.g., “Ala.”→“ ”), yielding the lowest SP-MCQA ACC despite lowest WER. These observations underscore the need for robust text normalization and key-information accuracy evaluation beyond conventional metrics.

5. CONCLUSION

In this work, we propose SP-MCQA, a framework that evaluates the accuracy of key information in synthesized speech beyond the word level for TTS models. SP-MCQA demonstrates that even SOTA TTS models exhibit critical weaknesses in key information when handling real-world speech complexity. Key findings include: (1) significant discrepancies between word-by-word accuracy and key information accuracy; (2) phonetic errors and text-normalization challenges in uncommon contexts and irregular patterns — particularly with names, numbers, and abbreviations — with each model exhibiting distinct error patterns, all of which should be addressed in future speech synthesis research. While our approach paves the way for high-level evaluation, limitations such as the substantial manual effort required for human evaluation remain. Future work will explore leveraging Audio LLMs for more efficient and scalable assessment and extend this work to other languages to provide clearer guidance for improving multilingual TTS systems.

6. REFERENCES

- [1] Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath, “VoiceCraft: Zero-shot speech editing and text-to-speech in the wild,” in *Proc. ACL*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar, Eds., Aug. 2024.
- [2] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al., “Seed-tts: A family of high-quality versatile speech generation models,” *arXiv preprint arXiv:2406.02430*, 2024.
- [3] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al., “Naturalspeech 3: zero-shot speech synthesis with factorized codec and diffusion models,” in *Proc. ICML*, 2024.
- [4] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karter, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al., “Voicebox: Text-guided multilingual universal speech generation at scale,” *NeurIPS*, vol. 36, 2024.
- [5] Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing, “Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis,” *arXiv preprint arXiv:2411.01156*, 2024.
- [6] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen, “F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching,” *arXiv preprint arXiv:2410.06885*, 2024.
- [7] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al., “Cosyvoice 2: Scalable streaming speech synthesis with large language models,” *arXiv preprint arXiv:2412.10117*, 2024.
- [8] Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu, “MaskGCT: Zero-shot text-to-speech with masked generative codec transformer,” in *Proc. ICLR*, 2025.
- [9] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.
- [10] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, “Libri-tts: A corpus derived from librispeech for text-to-speech,” in *Proc. Interspeech*, 2019.
- [11] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al., “Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio,” *arXiv preprint arXiv:2106.06909*, 2021.
- [12] Shang-Bao Luo, Hung-Shin Lee, Kuan-Yu Chen, and Hsin-Min Wang, “Spoken multiple-choice question answering using multimodal convolutional neural networks,” in *Proc. IEEE ASRU Workshop*, 2019.
- [13] Noussaiba Djeflal, Hamza Kheddar, Djamel Addou, Ahmed Cherif Mazari, and Yassine Himeur, “Automatic speech recognition with bert and ctc transformers: A review,” in *Proc. IC2EM*, 2023, vol. 1, pp. 1–8.
- [14] Vatsal Raina, Adian Liusie, and Mark Gales, “Analyzing multiple-choice reading and listening comprehension tests,” 2023.
- [15] Matthijs Westera, Laia Mayol, and Hannah Rohde, “TED-Q: TED talks and the questions they evoke,” in *Proc. LREC*, May 2020.
- [16] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al., “Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation,” in *Proc. IEEE SLT Workshop*, 2024, pp. 885–890.
- [17] Max Bain, Jaesung Huh, Tengda Han, and Andrew Senior, “Whisperx: Time-accurate speech transcription of long-form audio,” in *Proc. Interspeech*, 2023.
- [18] Hervé Bredin, “pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe,” in *Proc. Interspeech*, 2023, pp. 1983–1987.
- [19] Steven Bird and Edward Loper, “NLTK: The natural language toolkit,” in *Proc. ACL Demo*, July 2004.
- [20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al., “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.
- [21] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [22] Chandan KA Reddy, Vishak Gopal, and Ross Cutler, “Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *Proc. ICASSP*, 2022, pp. 886–890.