forum acusticum 2026
8-12 September · Graz, Austria

EVALUATION OF AI-BASED TEXT-TO-SPEECH GENERATORS

Steve Göring[1], Usman Khalid[1], Annika Neidhardt[2]

[1] *Audiovisual Technology Group, Technische Universität Ilmenau, Germany*
[2] *Audio Engineering, Faculty of Media, Mittweida University of Applied Sciences, Germany*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Text-to-speech systems are an essential building block in our modern AI-driven world. Various models have been established and are often also provided as open source. In this study, we evaluate ten different AI-based text-to-speech systems, with eight of them selected from the publicly available toolbox Coqui.ai TTS. In addition, with ChatGPT and Google Translate, two proprietary models are included for a wider evaluation. We conducted a perceptual evaluation based on ten pre-defined text-prompts. In an online study, participants rated audio quality, clarity, and naturalness. Our results indicate that the generators VITS and Google Translate are the best regarding audio quality and clarity, while ChatGPT achieves the highest score in naturalness. In addition, a high correlation between quality, clarity, and naturalness could be observed. We further carried out an objective evaluation with low- and high-level audio features, where the Meta Audiobox Aesthetics shows the highest correlation with all three dimensions. In the future, more generators and text-prompts should be included in similarly designed evaluation studies. Further, a comparison to real speech will be a valuable next step.

Keywords: *text-to-speech, artificial intelligence, quality*

1. Introduction

A wide range of artificial intelligence (AI) technologies is currently available, many of which can be leveraged to enhance both professional workflows and interaction with technical systems. One AI-technology is, for instance, text-to-speech (TTS) generation. TTS systems can be used as an interface to interact with computers. The early systems, for example eSpeak NG, were limited in their capabilities. Recent developments, such as generative AI and natural language processing, further advanced this technology and resulted in more sophisticated TTS models [1]. For example, Tortoise-TTS v2 is a text-to-speech system [2] allowing the synthesis of multiple distinct voices. Furthermore, VITS is an end-to-end model [3], skipping an intermediate representation as spectrogram. Those TTS systems are typically used in assistant systems, such as Alexa, Siri, Cortana, or OpenAI Whisper, in combination with a speech-to-text model [4]. The majority of TTS systems are only evaluated in small samples and often not compared in benchmarks for comparison. For this reason, we tackle in this paper the evaluation of different AI-based text-to-speech systems with human ratings and objective metrics.

We selected 10 different TTS generators, i.e., ChatGPT's web TTS, Google Translate's TTS and eight selected from the open source Coqui.ai TTS toolbox, eight female voices and two male voices were generated. The generators are taken as they are with no dedicated retraining. For the evaluation, we designed an online study, where participants rated audio quality, clarity, and naturalness. For the study, we selected 10 different text-prompts and thus, generated in total 100 stimuli to be rated by the participants.

The evaluation focuses on the human annotations and furthermore, we compare the ratings with state-of-the-art objective low- and high-level audio features. The results indicated that the TTS generator VITS, Google Translate, ChatGPT are the best with respect to audio quality, and clarity. While the generator ChatGPT is best for naturalness, followed by VITS and Tortoise-TTS v2, Google Translate shows a lower naturalness score in contrast to the high-rankings for audio quality and clarity. Thus, VITS or ChatGPT are the best TTS systems overall, considering all three attributes. Also, quality, clarity, and naturalness show quite high correlation, thus depending on each other. An analysis with available predictive models indicates, that the predictions of the Meta Audiobox Aesthetics [5] have the highest correlation with subjective annotations. The data, evaluation code, and annotations are published as open-source[1] for reproducibility.

***Corresponding author**: steve.goering@tu-ilmenau.de.*

[1] `https://github.com/avt-imt/text_to_speech`

The paper is structured as follows. In Sec. 2 a brief overview of the evaluation of TTS and TTS systems is provided. The subsequent Sec. 3 presents our selected generators and text-prompts and processing. Furthermore, in Sec. 4 the evaluation based on the human annotations is presented. Finally, the paper concludes in Sec. 5 with a summary and ideas for future work.

2. Related Work

TTS systems existed already before AI approaches took over. The early systems ranged, e.g., from manual mixing approaches, to automated systems (rule-based) converting text to speech [6]. However, automated systems still sounded synthetic and required adjustments. One of the first deep learning models was WaveNet [7], producing more realistic speech according to an evaluation based on spectrogram generation. Creating speech that satisfies the users' perceptual demands is the key goal. As a first step, suitable quality constructs needed to be identified. For transmitted speech, the quality can be decomposed into the perceptual dimensions noisiness, coloration and discontinuity [8], which forms the basis of the subjective assessment. For synthesized speech, Hinterleitner [9] proposed a test protocol for the efficient identification of those dimensions in a listening test and derived a set of five universal perceptual quality dimensions for TTS signals. These are *naturalness of voice*, *prosodic quality*, *fluency and intelligibility*, *absence of disturbances*, and *calmness*. Similarly, Seebauer et al. [10] propose a revised paradigm for meaningful evaluation of TTS and voice conversion (VC) systems based on two experiments. First, they determined descriptive terms by querying naïve listeners on their impressions of TTS and VC system. Second, they refined these terms into dimensions of quality and similarity by manual clustering. Besides *intelligibility* and *naturalness*, they identified *tempo* and *demographics*.

In listening tests, following standardized procedures is important to avoid inconsistencies or missing details [11]. Quality and naturalness were most common in the analyzed studies. However, clear instructions and reproducibility need to be maintained. Overall, speech technology evaluations are highly context-sensitive, as recent studies [12], [13] have demonstrated that contextualization and the quality dimensions being measured (e.g., naturalness versus appropriateness) significantly influence Mean Opinion Score (MOS) and similar evaluations of TTS systems. Context can be incorporated by integrating TTS into their intended real-world applications. The ITU recommendation P.85 highlights the importance of contextual framing when assessing TTS quality within specific applications. A study [13] found that evaluations conducted without framing showed minimal differentiation between modern TTS voices often leading to ceiling effects where all systems receive similarly high scores. However, introducing framing—either explicit or implicit—tended to lower the median MOS ratings even for state-of-the-art voices. This suggests that framing provides a more nuanced assessment by aligning user expectations with the specific use case highlighting the impact of context on perceived quality. Moreover, Mendelson and Aylett [14] evaluated TTS considering complex interactions with an avatar in comparison to traditional audio-only listening tests. The results indicate that the context and content may confound perception of voice quality.

The environment is another factor influencing quality perception. A TTS introduced in [15] aimed to incorporate acoustic environment aspects into the generation. Objective and subjective evaluation showed that the system can effectively synthesize speech that carries speaker and environment attributes.

Subjective assessment is limited by the number of stimuli and participants being recruited, however, objective prediction models, trained on human labels, can leverage a large-scaled evaluation. E.g., Mittag et al. [16] presented an objective prediction model for synthetic speech naturalness to evaluate Text-To-Speech or Voice Conversion systems. The model is trained end-to-end and includes self-attention mechanism. In addition to overall speech quality, the model predicts noisiness, coloration, discontinuity, and loudness.

Several TTS systems are accessible as open-source, however the majority of studies focuses on individual TTS system evaluation and a cross comparison of different established TTS approaches is missing, which is the aim in our paper.

3. Approach

In the following we briefly describe the selection of the to-be-evaluated text-to-speech generators, followed by the crafted text-prompts and the unification required for the subjective evaluation.

3.1. Text-to-speech Generators

Various TTS generators are established. We selected ten TTS models, eight are included in the Coqui TTS project [1]. The Coqui TTS project is an open-source library for TTS synthesis including > 34 models. For the evaluation we selected, Tacotron2 [17], Glow-TTS [18], SpeedySpeech [19], VITS [3], Fastpitch [20], Overflow [21], Neural HMM TTS [22], and Tortoise-TTS v2 [2] (male).

This list was further extended by two API only TTS services, namely Google Translate (english), and ChatGPT web interface (male) to compare the open source TTS approaches with proprietary standards. For Google translate, and ChatGPT, we extracted the generated output of the webpage using web-scraping techniques. Except ChatGPT and Tortoise-TTS v2 all other voices are female.

3.2. Text prompts

Overall, we selected 10 text-prompts, listed in Table 1. The word types have been extracted using spaCy [23] and the phonemes with the CMU Pronouncing Dictionary included in NLTK [24].

The text-prompts tp0 and tp1 (class *D*), selected from [25], provide moderate phoneme diversity to

Table 1: Selected text-prompts and characterization.

id	class	prompt	unique words	unique phonemes	syllables	nouns	verbs	adjectives	phonemes per word
tp0	*D*	Nobody could've imagined that it'd become a breakout hit.	9	21	15	2	2	0	4.000
tp1	*D*	Nowadays, a voice can be cloned with mere minutes of data.	11	23	15	3	1	1	3.455
tp2	*I*	I ain't got no money.	5	9	6	1	1	0	2.200
tp3	*I*	She don't like vegetables.	4	14	6	1	1	0	4.000
tp4	*E*	The sky was ablaze with the colors of the sunset, painting the horizon in hues of orange and pink.	16	28	26	6	1	2	3.526
tp5	*E*	With a heavy heart, he watched her walk away, the weight of unspoken words hanging in the air.	17	28	23	4	3	2	3.111
tp6	*E*	His words were like daggers, each one piercing deeper than the last.	12	24	15	3	1	2	3.417
tp7	*I*	Between you and I, this is a bad idea.	9	17	12	1	0	1	2.778
tp8	*E*	The garden was a riot of colors, with flowers in full bloom swaying gently in the summer breeze.	17	31	25	7	1	1	3.500
tp9	*E*	The thunderstorm raged outside, its fury echoing the turmoil within her heart.	12	32	20	4	2	0	4.333

challenge the TTS systems. For example, tp1 consists of 11 words (3 nouns, 1 verb phase, 1 adjectives), 23 unique phonemes, and ≈ 3.46 phonemes per word. Structurally tp1 consists of 15 syllables ensuring rhythmic complexity that tests prosodic capabilities. The technical vocabulary such as "cloned" and "minutes" adds to the challenge by requiring precise articulation of less common technical terms. The prompts tp2, tp3, tp7 (class *I*) form a category of sentences which are grammatically incorrect but used very frequently in everyday life. These sentences were selected to analyze TTS system ability to handle informal grammar and maintain authenticity. Errors like missing articles or verb conjugation issues may confuse phoneme prediction leading to unnatural pronunciation [26].

Furthermore, to evoke emotions and create vivid mental images making the listener feel more connected to the experience or sentiment being described, we included tp4, tp5, tp6, tp8, and tp9 (class *E*). All text prompts, except tp0, tp1, were defined using different ChatGPT prompts, to finetune them, e.g. for class *I* prompts, we used "Examples of sentences that are grammatically incorrect but used frequently in everyday life", and class *E* prompts were identified using "English sentences with emotions and vivid mental images" as text-prompt for ChatGPT.

3.3. Stimuli preparation

Because the different TTS systems provide different output formats (WAV, AAC or MP3), we converted the stimuli to FLAC [27] as unification. Only Google translate (MP3) and ChatGPT (AAC) used lossy compression, however the internal processing of the included TTS systems may also be lossy. The unification to FLAC further allows easy inclusion in our web-based test framework.

In addition to the format unification, we analyzed the loudness levels of the generated audio. It was required to perform a loudness normalization to unify the listening experience. The normalization was done using a library pyloudnorm [28] which provides an implementation of ITU-R BS.1770-4 with -14.0 dB.

4. Evaluation

After the stimuli were created, processed and unified, we designed an online study using our publicly available tool AVrateVoyager [29]. Adjustments similar to [30] were performed, covering an additional step for the loudness calibration of the participants with a given example. The participants were instructed to rate naturalness, quality, and clarity. Naturalness was introduced as the closeness with human voice (1=very far away, 5=very close), quality (1-5; 1=very low, 5=very high): technical audio quality, and clarity (1-5; 1=very bad, 5=very good): quality of being easily understandable. These definitions were also visible for each stimulus as a hint in the web-interface. As rating scheme we used Absolute Category Rating.
Overall, the processing resulted in 100 different stimuli with approximately 3 s-4 s duration each. These stimuli were rated by 14 valid users (10 male, 4 female). More participants (25 recruited from Clickworkers) took part in the study, however, we excluded participants with incomplete ratings, low English level skills, participants reporting the need of hearing aids, and users identified as outliers based on the detection described in ITU-T Recommendation P.913. The overall duration of the test was approximately 30 min. The majority of users were in the age range 30-39, and 40-49 with each category covered by 4 participants.

4.1. Subjective analysis of quality, clarity, and naturalness

We calculated for each stimulus the mean rating considering quality, clarity, and naturalness.
Fig. 1 shows the rating distributions for naturalness, quality and clarity for each text-prompt. Overall, all

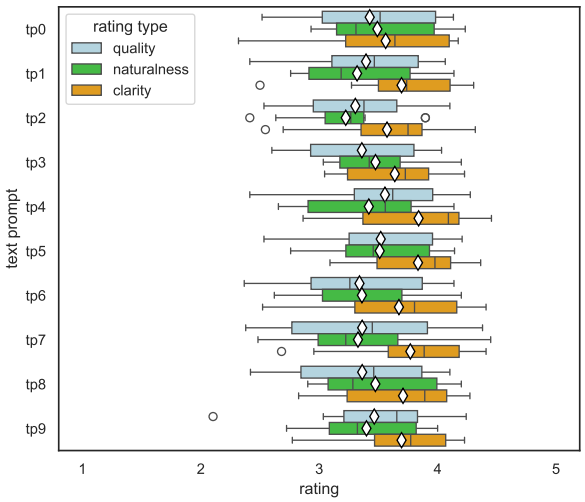


Figure 1: Ratings grouped by text-prompt.

text-prompts are in the higher quality, naturalism and clarity range. The highest mean quality rating has tp4 (3.56), the lowest tp2 (3.31). For naturalness, the lowest mean scores is for tp2 (3.23), and the highest for tp5 (3.51). Clarity has the highest mean score for tp4 (3.84), and the lowest for tp0 (3.56).

In addition, in Fig. 2 the ratings for quality, naturalness, and clarity regarding the used TTS system are shown. We further performed paired t-tests for side-by-side ranked TTS systems, non-significant results are not visualized.

Fig. 2a provides an overview of the quality ratings. The models VITS (overall best model) and Google_Translate have the highest scores. The lowest scores are for neural_hmm, SpeedySpeech and Tacotron2. In general, the best 3 models are in a similar high range. Furthermore, Fig. 2b compares the ratings of naturalness for each of the models. Similarly to the quality evaluation VITS is highly ranked at place 2, only ChatGPT provides a significant higher naturalness rating. The least natural rated TTS systems are SpeedySpeech, neural_hmm, and Tacotron2, which is align with their low quality ratings. As last, we evaluated clarity, visualized in Fig. 2c. Here, VITS, Google_Translate, ChatGPT have the highest scores, with no significant differences. The lowest clarity scores are for neural_hmm, Tacotron2, and SpeedySpeech, similar to quality and naturalness. In general, the TTS systems, FastPitch, Tortoise-v2, Glow-TTS and overflow are always ranked in between the top-3 and worst-3 for all three dimensions.

Further, we analyzed how the different rated dimensions are connected. Tab. 2 provides the correlation values. The dimensions quality vs. clarity are highly correlated, followed by clarity vs. naturalness and quality vs. naturalness. Thus an interlink between the different dimensions can be observed, however only quality and clarity are highly connected.

Table 2: Correlation values; rounded to 3 decimals.

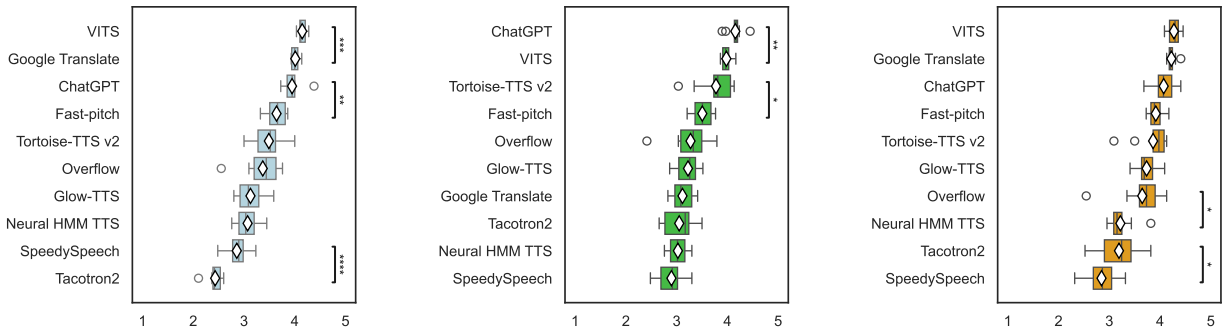
Comparison	Pearson	Kendall	Spearman
quality vs. clarity	0.861	0.733	0.904
clarity vs. naturalness	0.712	0.532	0.683
quality vs. naturalness	0.688	0.506	0.674

4.2. Analysis of audio models for quality, clarity, and naturalness

First, we checked low-level features (LLF), like spectral or temporal features: zero-crossing rate (zcr), speaking_rate, frequency bands (f1, f2, f3), energy of the signal, spectral centroid/bandwidth/flatness, speech-to-reverberation modulation energy ratio (srmr) using librosa [31], srmrpy [32] and parselmouth [33], compare Tab 3. Because LLF showed no strong correlation with the subjective ratings, we focused on high-level features (HLF) which are part of prediction models, e.g., covering speech quality, i.e, MOSNet [34], NISQA [16] (overall quality=nisqa_q, naturalness=nisqa_n [35]), WVMOS [36], [37], DNSMOS [38], Meta Audiobox Aesthetics (MAE) [5] (Meta-CE=Content Enjoyment, Meta-CU=Content Usefulness, Meta-PC=Production Complexity, Meta-PQ=Production Quality) as SoA models for the analysis. Tab. 3 summarizes the results. The best models for quality, clarity, and naturalness are part of MAE (Meta-CU, Meta-PQ, and Meta-CE). E.g. Meta-CU has a 0.762 Pearson correlation for quality, a 0.649 Pearson correlation for clarity, and 0.472 for naturalness. Overall, naturalness is the dimension which is less correlated with objective models, making the development of models for naturalness more relevant. Moreover, the NISQA model predictions (nisqa_q, nisqa_n), and DNSMOS have medium correlations or lower with the analyzed dimensions. The least performing models are MOSNet, Meta-PC, and WVMOS across the three dimensions.

5. Conclusion and Future Work

This study compares a selection of TTS-systems regarding the resulting quality, naturalness and clarity of the synthesized speech. The evaluation is based on ten TTS systems and ten different text prompts. For the subjective evaluation, perceptual ratings were collected in an online study using AVrateVoyager. The results indicate that the best models for quality, naturalness, and clarity are VITS, ChatGPT, and Google Translate, with Google Translate exhibiting a lower ranked naturalness. An objective evaluation using high-level prediction models shows that models can predict quality, clarity, and naturalness. However, the models have only a medium-to-low correlation with the subjective scores, indicating the need for the development and advancement of prediction models for those perceptual dimensions. An evaluation comparing with real voices as reference is still pending. Furthermore, a study with a broader range of TTS-systems, voice cloning or more participants will be of interest for future work.



(a) quality **(b)** naturalness **(c)** clarity

Figure 2: Perceptual ratings for TTS, ordered by mean rating per TTS.

6. Acknowledgments

We thank the "AG Wissenschaftliches Rechnen" of the TU Ilmenau for providing computing resources.

References

[1] G. Eren and T. C. T. Team, *Coqui tts*, version 1.4, A deep learning toolkit for Text-to-Speech, battle-tested in research and production, 2021. [Online]. Available: `https://github.com/coqui-ai/TTS`.

[2] J. Betker, "Better speech synthesis through scaling," *arXiv preprint arXiv:2305.07243*, 2023.

[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Int. conf. on machine learning*, PMLR, 2021, pp. 5530–5540.

[4] H. Ahlawat, N. Aggarwal, and D. Gupta, "Automatic speech recognition: A survey of deep learning techniques and approaches," *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 201–237, 2025, ISSN: 2666-3074. DOI: `10.1016/j.ijcce.2024.12.007`.

[5] A. Tjandra et al., "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," 2025. [Online]. Available: `https://arxiv.org/abs/2502.05139`.

[6] B. Lindquist, "The art of text-to-speech," *Critical Inquiry*, vol. 50, no. 2, pp. 225–251, 2024.

[7] A. Van Den Oord et al., "Wavenet: A generative model for raw audio," *arXiv preprint:1609.03499*, vol. 12, no. 1, 2016.

[8] M. Wältermann, *Dimension-based quality modeling of transmitted speech*. Springer, Berlin, 2013. DOI: `10.1007/978-3-642-35019-1`.

[9] F. Hinterleitner, *Quality of Synthetic Speech*. Springer Singapore, 2017. DOI: `10.1007/978-981-10-3734-4`.

[10] F. M. Seebauer, M. Kuhlmann, R. Haeb-Umbach, and P. Wagner, "Discerning dimensions of quality for state of the art synthetic speech," in *Proceedings of the 20th International Congress of Phonetic Sciences*, 2023.

[11] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely, and J. Gustafson, "Stuck in the mos pit: A critical analysis of mos test methodology in tts evaluation," in *12th Speech Synthesis Workshop (SSW) 2023*, 2023.

[12] J. O'Mahony, P. O. Gallegos, C. Lai, and S. King, "Factors affecting the evaluation of synthetic speech in context.," in *SSW*, 2021, pp. 148–153.

[13] J. Edlund, C. Tånnander, S. Le Maguer, and P. Wagner, "Assessing the impact of contextual framing on subjective tts quality," in *Interspeech*, ISCA-International Speech Communication Association, 2024, pp. 1205–1209.

[14] J. Mendelson and M. P. Aylett, "Beyond the listening test: An interactive approach to tts evaluation.," in *INTERSPEECH*, Stockholm, 2017, pp. 249–253.

[15] D. Tan, G. Zhang, and T. Lee, "Environment aware text-to-speech synthesis," in *Proc. Interspeech 2022, 481-485*, 2022.

[16] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *arXiv preprint arXiv:2104.09494*, 2021.

[17] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[18] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[19] J. Vainer and O. Dušek, "SpeedySpeech: Efficient neural speech synthesis," *arXiv preprint arXiv:2008.03802*, 2020.

[20] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *ICASSP*, 2021, pp. 6588–6592.

Table 3: Correlation of ratings and features; 3 decimals, sorted by Pearson per rating type group; M_ = mean.

Rating	Feature	Type	Pearson	Kendall	Spearman
quality	Meta-CU	HLF	0.762	0.526	0.727
	Meta-PQ	HLF	0.749	0.523	0.708
	Meta-CE	HLF	0.587	0.386	0.562
	nisqa_q	HLF	0.569	0.393	0.562
	speaking_rate	LLF	0.466	0.316	0.450
	dnsmos	HLF	0.422	0.245	0.357
	nisqa_n	HLF	0.347	0.220	0.316
	M_energy	LLF	0.224	0.148	0.228
	rms_energy	LLF	0.220	0.148	0.228
	wvmos	HLF	0.178	0.105	0.184
	mosnet	HLF	0.155	0.155	0.220
	M_bandwidth	LLF	0.123	0.043	0.077
	M_centroid	LLF	0.037	0.015	0.028
	srmr_score	LLF	0.010	0.022	0.034
	M_f0	LLF	-0.022	0.022	0.027
	M_zcr	LLF	-0.077	-0.046	-0.069
	M_f1	LLF	-0.298	-0.158	-0.236
	M_f2	LLF	-0.454	-0.272	-0.404
	M_f3	LLF	-0.494	-0.385	-0.545
	M_flatness	LLF	-0.509	-0.387	-0.535
	Meta-PC	HLF	-0.512	-0.354	-0.522
clarity	Meta-CU	HLF	0.649	0.479	0.682
	Meta-PQ	HLF	0.572	0.430	0.623
	Meta-CE	HLF	0.496	0.342	0.513
	nisqa_q	HLF	0.474	0.353	0.513
	speaking_rate	LLF	0.448	0.320	0.472
	nisqa_n	HLF	0.359	0.260	0.373
	dnsmos	HLF	0.283	0.201	0.303
	mosnet	HLF	0.254	0.225	0.336
	M_energy	LLF	0.242	0.162	0.243
	rms_energy	LLF	0.239	0.162	0.243
	M_bandwidth	LLF	0.195	0.126	0.188
	wvmos	HLF	0.155	0.102	0.165
	M_centroid	LLF	0.134	0.080	0.112
	M_f0	LLF	0.077	0.060	0.090
	M_zcr	LLF	0.006	-0.010	-0.029
	srmr_score	LLF	-0.071	-0.011	0.006
	M_f1	LLF	-0.196	-0.110	-0.169
	M_f2	LLF	-0.339	-0.207	-0.315
	M_f3	LLF	-0.411	-0.324	-0.496
	M_flatness	LLF	-0.428	-0.354	-0.520
	Meta-PC	HLF	-0.501	-0.376	-0.561
naturalness	Meta-CE	HLF	0.504	0.334	0.489
	Meta-CU	HLF	0.472	0.308	0.447
	Meta-PQ	HLF	0.411	0.270	0.392
	speaking_rate	LLF	0.328	0.211	0.311
	M_f0	LLF	0.246	0.184	0.266
	nisqa_q	HLF	0.245	0.161	0.248
	M_energy	LLF	0.210	0.150	0.221
	rms_energy	LLF	0.201	0.150	0.221
	dnsmos	HLF	0.120	0.066	0.094
	nisqa_n	HLF	0.112	0.111	0.176
	mosnet	HLF	0.096	0.118	0.173
	M_zcr	LLF	-0.090	-0.045	-0.073
	wvmos	HLF	-0.164	-0.111	-0.165
	M_f1	LLF	-0.176	-0.119	-0.180
	M_centroid	LLF	-0.208	-0.126	-0.191
	Meta-PC	HLF	-0.231	-0.160	-0.246
	M_bandwidth	LLF	-0.252	-0.155	-0.232
	M_flatness	LLF	-0.328	-0.256	-0.368
	M_f3	LLF	-0.395	-0.271	-0.391
	M_f2	LLF	-0.444	-0.295	-0.433
	srmr_score	LLF	-0.509	-0.321	-0.441

[21] S. Mehta, A. Kirkland, H. Lameris, J. Beskow, É. Székely, and G. E. Henter, "Overflow: Putting flows on top of neural transducers for better tts," *arXiv preprint arXiv:2211.06892*, 2022.

[22] S. Mehta, É. Székely, J. Beskow, and G. E. Henter, "Neural hmms are all you need (for high-quality attention-free tts)," in *ICASSP*, 2022, pp. 7457–7461.

[23] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "Spacy: Industrial-strength natural language processing in python," 2020.

[24] S. Bird, E. Klein, and E. Loper, *NLP with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[25] hecko-yes. "Tts-prompts." `https://github.com/hecko-yes/tts-dataset-prompts`.

[26] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *speech communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[27] M. van Beurden and A. Weaver, "Free lossless audio codec," *IETF Datatracker*, 2022.

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

[29] S. Göring, R. Rao Ramachandra Rao, S. Fremerey, and A. Raake, "AVRate Voyager: an open source online testing platform," in *Int. Wksp. on Multimedia Signal Processing*, 2021, pp. 1–6.

[30] L. Treybig, F. Klein, S. Werner, S. V. A. Garí, and S. Göring, "Predicting the perceptibility of room acoustic variations using generalized linear mixed models," *J. Audio Eng. Soc. 73(11)*, 2025.

[31] B. McFee et al., "Librosa: Audio and music signal analysis in python.," 2015.

[32] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[33] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," 2018.

[34] C.-C. Lo et al., "Mosnet: Deep learning based objective assessment for voice conversion," in *Proc. Interspeech 2019*, 2019.

[35] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," *arXiv preprint arXiv:2104.11673*, 2021.

[36] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: A unified framework for bandwidth extension and speech enhancement," in *IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[37] P. Andreev. "WV-MOS Github Project." `https://github.com/AndreevP/wvmos`.

[38] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2022.