The VoiceMOS Challenge 2022

Wen-Chin Huang[1†], *Erica Cooper*[2†], *Yu Tsao*[3], *Hsin-Min Wang*[3], *Tomoki Toda*[1], *Junichi Yamagishi*[2]

[1]Nagoya University, Japan
[2]National Institute of Informatics, Japan
[3]Academia Sinica, Taiwan

voicemos2022@nii.ac.jp

Abstract

We present the first edition of the VoiceMOS Challenge, a scientific event that aims to promote the study of automatic prediction of the mean opinion score (MOS) of synthetic speech. This challenge drew 22 participating teams from academia and industry who tried a variety of approaches to tackle the problem of predicting human ratings of synthesized speech. The listening test data for the main track of the challenge consisted of samples from 187 different text-to-speech and voice conversion systems spanning over a decade of research, and the out-of-domain track consisted of data from more recent systems rated in a separate listening test. Results of the challenge show the effectiveness of fine-tuning self-supervised speech models for the MOS prediction task, as well as the difficulty of predicting MOS ratings for unseen speakers and listeners, and for unseen systems in the out-of-domain setting.

Index Terms: VoiceMOS Challenge, synthetic speech evaluation, mean opinion score, automatic speech quality prediction

1. Introduction

Speech synthesis technologies such as text-to-speech synthesis (TTS) and voice conversion (VC) are a very active area of research, and this field relies on judgments of quality of the synthesized speech. The gold standard for evaluation is listening tests, where human raters listen to samples generated by different synthesis methods and give their opinions. One popular type of listening test is Mean Opinion Score (MOS) [1], in which listeners are presented with synthesized samples one by one, and asked to rate some aspect of the speech, such as how natural it sounds, on a Likert scale, typically from 1-5.

With the popularity of crowdsourcing platforms, conducting listening tests has become easier than in the past, when researchers would have to recruit and schedule participants locally to come into the lab to listen to samples in person. MOS tests also enable the implicit comparison of many different synthesis systems – tests such as A/B testing, which ask listeners to directly compare samples from two different systems, quickly become impractical as the number of systems increases, since all pairs of systems must be compared. Nevertheless, listening tests in general and MOS tests in particular are not without their drawbacks. Listening tests are time-consuming and costly, and they cannot be used as a loss term in a training objective when developing synthesis systems. Furthermore, listening tests are heavily context-dependent. Each listening test gathers opinions from a different set of listeners, and asks them to consider different sets of systems, whose overall range of quality may be very different. The lexical content of the synthesized samples and even the instructions given to participants also differ from test to test. This means that although MOS testing results in numerical values representing the quality of each system, these numbers cannot be meaningfully compared across different listening tests.

Intrusive speech quality metrics such as PESQ [2] are not suitable for evaluating synthesized speech because they require matching ground-truth reference audio. Synthesized speech may have different prosody from a ground-truth reference, unlike natural speech processed by a codec or in noisy conditions, for which these metrics were designed. Given that MOS tests result in a set of audio samples paired with their ratings from human listeners, there is no surprise that researchers have begun to apply data-driven, machine-learning-based approaches to develop automatic non-intrusive MOS prediction models [3–13]. However, given the context-dependency of MOS ratings, data from different listening tests cannot be combined, and so these models are typically trained on data from one listening test that contains samples from a limited range of types of synthesis models. While these MOS prediction models can perform well for samples from the same listening test, they typically fail to generalize well to other listening test contexts or even to unseen synthesis methods from the same listening test [14].

In order for MOS prediction models to be a useful tool for speech synthesis researchers, these issues need to be addressed and more improvement is needed in the state of the art in the MOS prediction task. With these aims, we developed the **VoiceMOS Challenge,** a shared task using common datasets for MOS prediction. We conducted a large-scale MOS listening test to cover many different types of TTS and VC systems spanning many years [14,15], and this dataset forms the basis of the challenge. We also introduced an "out-of-domain" track to focus research efforts on how to adapt MOS prediction models to new listening test contexts. In the following, we describe the organization of the challenge and present some analysis of the predictions made by participants' MOS prediction systems.

2. Challenge Description

The challenge was hosted on CodaLab[1], an open-source web-based platform for reproducible machine learning research. Below we introduce the dataset used in each track. A summary is in Table 1. We also describe the challenge phases, evaluation metrics, and baseline prediction systems.

2.1. Tracks and Datasets

2.1.1. Main track

The data for the main track of the challenge comes from a large-scale listening test covering many types of TTS and VC sys-

[†]Equal contribution.

[1]https://codalab.lisn.upsaclay.fr/competitions/695

Table 1: *Summary of the main track and out-of-domain (OOD) track datasets.*

Track	Lang	# Samples			# ratings per sample
		Train	Dev	Test	
Main	Eng	4,974	1,066	1,066	8
OOD	Chi	Label: 136 Unlabel: 540	136	540	10-17

tems that we conducted in our prior work [15]. The listening test consists of 38 samples from each of 187 different synthesis systems including natural speech, and each sample was rated by 8 different listeners. The samples come from many past years of Blizzard Challenges (BC) [16–21] and Voice Conversion Challenges (VCC) [22–24], as well as published samples from ESPnet-TTS [25]. These samples cover a wide range of different types of synthesis systems and methods spanning many years. Listeners rated each sample for naturalness on a discrete, 1-incremented scale from 1 (very bad) to 5 (very good). We also collected demographics about listeners such as gender, age range, and whether or not they have any hearing impairment. We used the training, development, and test splits from our prior work [14] which followed a 70%/15%/15% partition, and which were chosen to hold out some unseen synthesis systems, speakers, texts, and listeners in the development and test sets while matching the overall distributions of ratings as closely as possible between the sets.

2.1.2. *Out-of-domain track*

We define "out-of-domain" (OOD) as data that is from a different listening test. For our OOD track, we chose the Blizzard Challenge 2019 listening test data [26], which was not included in the main track data. In addition to this data originating from a different listening test with different synthesis systems and listeners from the main track, this dataset consists of Chinese TTS samples whereas all of the samples in the main track dataset are in English, which presents a challenging language mismatch in the domain as well. The OOD data was split into training/unlabeled/development/test sets of 10%/40%/10%/40%. The unlabeled dataset was audio samples only without MOS ratings, provided to participants during the training phase in order to encourage experimentation with approaches such as semi-supervised learning. The small training set size was chosen to reflect a scenario where a small amount of labeled target-domain data is available, e.g., from a small pilot listening test. The main track data was allowed to be used in the OOD track in addition to the OOD data. We have made the scripts for obtaining the challenge data publicly available.[2]

2.2. Phases

The challenge was divided into four phases: the training phase, evaluation phase, break phase, and post-challenge.

2.2.1. *Training phase*

The training phase started on December 5, 2021, and ended on February 21, 2022. During these 11 weeks, participants had access to audio samples and their MOS labels and listener info for the training and development data (as well as the unlabeled audio set for the OOD track), and the purpose of this phase was for participants to develop their prediction systems. Participants could submit their MOS predictions for the development set to the CodaLab leaderboard, where teams' scores were publicly displayed. Participants were permitted to make up to 30 submissions to the leaderboard per day.

2.2.2. *Evaluation phase*

The evaluation phase was one week long, starting on February 21, 2022 and ending on February 28. Test set audio samples were released at the start of this phase. Participants did not have access to any MOS labels or listener info for the test set. Teams were permitted to make up to three submissions during this phase with their MOS predictions for the test set audio.

2.2.3. *Break phase and post-challenge phase*

We defined a 1-week break phase after the end of the evaluation phase, during which we froze new submissions so that we could conduct post-challenge analysis. After that was the post-challenge phase, during which we re-opened the leaderboard so that participants could continue making submissions.

2.3. Evaluation Metrics

Following [8], the evaluation criteria are system-level and utterance-level mean squared error (MSE), Linear Correlation Coefficient (LCC), Spearman Rank Correlation Coefficient (SRCC), and Kendall Tau Rank Correlation (KTAU). The reason to use multiple metrics is to assess the model performance via various aspects. Depending on the application, different metrics have their own respective roles. In scientific challenges like BC or VCC, we are more interested in the ranking of synthesis systems, so metrics like SRCC or KTAU are preferable. On the other hand, when it comes to assessing TTS/VC models under development, MSE is a more straightforward metric.

We distributed an evaluation script to avoid confusion and inconsistency. Due to the space limit, we only displayed utterance-/system-level MSE/SRCC on the leaderboard. Following [13], we used system-level SRCC as the primary metric for determining the ranking on the leaderboard.

2.4. Baseline Prediction Systems

We provided three open-source baseline MOS prediction systems to participants: SSL-MOS[3] [14], MOSA-Net[4] [27], and LDNet[5] [28], which correspond to team IDs B01, B02 and B03, respectively. SSL-MOS adds a simple linear fine-tuning layer for the MOS prediction task onto Fairseq[6] self supervised learning (SSL)-based models for speech, and in particular, the baseline is fine-tuned from Wav2vec 2.0 Base [29], which was trained on the LibriSpeech [30] corpus. MOSA-Net uses cross-domain features such as spectral information, complex features, raw waveform data, and features extracted from SSL models to estimate multiple speech assessment metrics simultaneously. Originally developed for noisy speech quality assessment, MOSA-Net serves as a good example to demonstrate

[2] https://doi.org/10.5281/zenodo.6572573

[3] https://github.com/nii-yamagishilab/mos-finetune-ssl

[4] https://github.com/dhimasryan/MOSA-Net-Cross-Domain

[5] https://github.com/unilight/LDNet

[6] https://github.com/pytorch/fairseq

how adapting such a prediction system to this task is possible. LDNet does not rely on any external data, but its specialized model structure and inference method allow for utilizing individual ratings to facilitate listener-dependent modeling.

3. Participants and Submitted Systems

In total, 22 teams made an evaluation phase submission, where 14 are from academia, five are from industry, two are joint academia-industry teams, and one is personal. Participants span from countries in Asia such as Japan, Taiwan, and China, to those in Europe, including the UK, Ireland, the Netherlands, Hungary, Romania, and Czechia. Among them, 21 and 15 teams participated in the main and OOD tracks, respectively.

During the training phase, a total of 223 submissions were made, as some teams frequently uploaded results to compete with each other. One of the teams, T17, made 33 submissions and turned out to be one of the top teams.

4. Results, Discussion and Analysis

The evaluation results of the main and OOD tracks are shown in Figures 1 and 2. Full numerical results for all metrics can be found in our supplementary materials online[7].

4.1. Comparison of Baseline Systems

Among the three baselines, B01 was the strongest in terms of system-level MSE and SRCC for the main track and system-level SRCC for the OOD track. B02 topped the system level MSE for the OOD track, and placed second among the other metrics. B03 was consistently the worst among the three baselines. As both B01 and B02 used SSL, it is shown that using extra data is helpful even if MOS ratings are absent.

For the main track, B01 had a system-level MSE and SRCC of 0.148 and 0.921, ranking 18th and 12th in the two metrics, respectively. The top prediction systems in system-level MSE and SRCC scored 0.090 and 0.939, respectively. As for the OOD track, B02 had a system-level MSE of 0.071, and B01 had a system-level SRCC of 0.975. The top prediction systems had an MSE of 0.030 and a SRCC of 0.979, respectively. This suggests that the gap between the baseline and the top prediction systems was not large. As one of the points of feedback from the participants suggests, the baseline was already strong and it was hard to obtain substantial improvements over it.

4.2. Analysis of Top Prediction Systems

The top-ranking teams in the main track all use methods that involve fine-tuning SSL models (either using our SSL-MOS baseline code or not). In fact, for the main track we can see from Figure 1 that teams are clearly arranged with SSL-finetuning-based approaches scoring the highest, models that use features extracted from SSL models without finetuning in the middle, and methods that do not use SSL ranking last. However, for the OOD track this arrangement does not hold true, with some teams that did not finetune SSL models outranking ones that did. Some popular modeling approaches that we observed in both the top prediction systems and the teams overall are ensembling, multi-task learning, and use of speech recognizers.

Seven teams (and 3 of the top 5 teams) made use of per-listener ratings, either by using the LDNet baseline or by implementing a similar method. No teams made use of the listener demographics that were provided, but one team made use of "listener group," since the structure of the main-track listening test had listeners grouped by the set of utterances that they rated.

For the OOD track, only three teams tried making use of the unlabeled data. T17 conducted their own listening test to get labels for the unlabeled data, and this strategy was very successful as they ranked first for system-level SRCC in the OOD track. Other strategies were to use the unlabeled data for task-adaptive pretraining [31], and to use trained models to "self-label" the unlabeled data. System-level SRCCs were overall higher in the OOD phase than in the main phase, but the MSEs also spanned a wider range; this indicates that although relative rankings of synthesis systems can be well-learned by pretraining on a larger dataset, predicting the precise scores for a different listening test with very little data from that test remains a challenge.

4.3. Sources of Difficulty for MOS Prediction

4.3.1. Seen vs. unseen categories

Similar to our prior work [14], we analyzed the difficulty of seen vs. unseen categories such as synthesis systems, speakers, and listeners. For each team, we gathered the squared errors for all utterances from e.g. seen speakers, and then for unseen speakers, and then conducted a two-sided t-test between these lists of squared errors to determine whether the unseen-category errors were significantly different (either more difficult or less difficult) from the seen ones at a level of $p \leq 0.05$.

Surprisingly, for the main track, we found that no teams found unseen synthesis systems to be significantly more difficult by this measure, except for B01 (as reported in [14]). Unseen listeners were significantly more difficult for 16 teams and 1 baseline (and significantly easier for 1 team). There was only one unseen speaker in the test set, but unseen speaker data was significantly more difficult for 6 teams and 1 baseline.

For the OOD track, unlike the main track, unseen synthesis systems were more difficult than seen ones in the OOD test set for 4 teams and 2 baselines. Unseen listeners were not significantly harder to predict for any teams. There are no unseen speakers in the OOD track test set.

4.3.2. Synthesis systems that are difficult to predict

Since unseen synthesis systems were not especially difficult to predict in the main track, we next tried to identify whether there were any particular synthesis systems that are consistently difficult for all teams, and to identify what makes them difficult. For all teams including baselines, we ranked synthesis systems based on system-level MSE, and found each team's "worst-10" synthesis systems. Then, for each synthesis system that appears in *any* "worst-10" list, we count how many teams have that system in their worst-10 list. The top 5 most difficult synthesis systems, judged as those that appeared in the most teams' worst-10 lists, are in Table 2.

Table 2: *Top 5 most difficult synthesis systems to predict in the main track, and information about them: System name, how many teams had it in their worst-10 list, ground-truth MOS, and standard deviation of MOS.*

Name	# Teams	MOS	Sdev. MOS
BC2009-T	22	2.65	0.88
BC2016-F	19	2.52	1.00
BC2008-B	15	3.22	0.98
BC2008-I	13	2.40	1.02
VCC2020-T10	12	3.88	0.83

[7] https://arxiv.org/abs/2203.11389

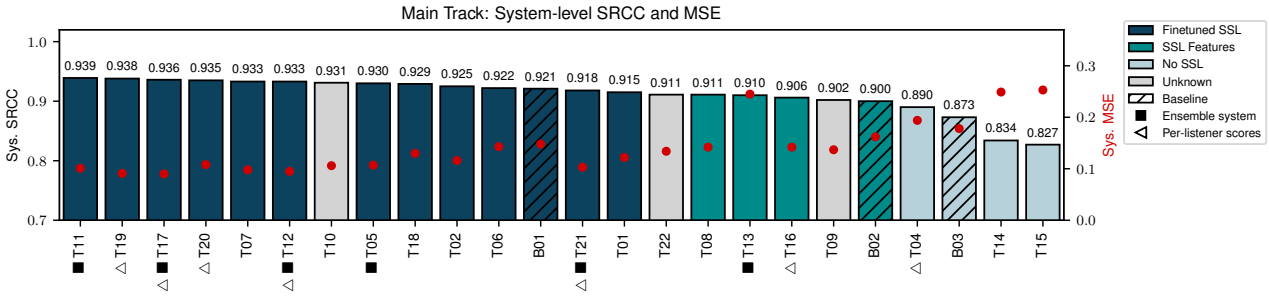


Figure 1: *Main track results*

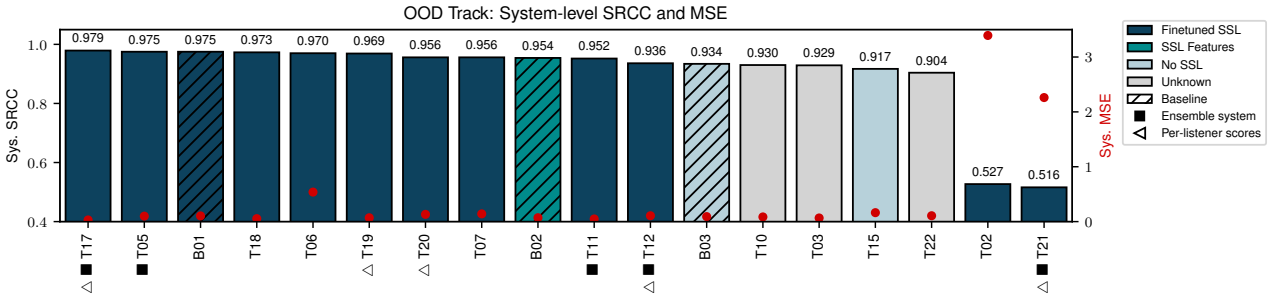


Figure 2: *OOD track results*

Table 3: *Linear correlation coefficients between system-level metrics, using the main track results.*

	MSE	LCC	SRCC	KTAU
MSE	1.00	-.875	-.862	-.870
LCC	-	1.00	.997	.994
SRCC	-	-	1.00	.994
KTAU	-	-	-	1.00

We did not find that factors such as type of synthesis or standard deviations of synthesis systems' ratings could explain their difficulty. We also considered that differences in the distributions of scores for each synthesis system's training and testing splits may affect difficulty. Although our training, development, and testing splits were carefully chosen to have well-matched distributions of MOS ratings, this was done for the overall data and not on a finer-grained system-by-system level. We measured the difference in the distributions of the splits for each synthesis system in the following manner: For every synthesis system in our main track data, we computed the earth-mover's distance (EMD) between the distribution of the ratings in the training set compared to the test set. System VCC2020-T10 had no samples in the training set since it was held out as an unseen system for the development set, so it was excluded from this analysis. We found that all four of the remaining difficult systems fall within the 15% of systems with the highest EMDs. Thus, unsurprisingly, we can conclude that a discrepancy in the training and test set distributions can adversely affect prediction ability; this is a well-understood phenomenon in machine learning.

4.4. Analysis of Metrics

We observed that ranking top at one metric does not guarantee superiority at other ones. For example, T11 ranked top in system-level LCC, SRCC and KTAU in the main track, but ranked 5th in system-level MSE. To understand the tendencies of each team at each metric, we calculated the linear correlation coefficients between all pairs of metrics using the main track results, as shown in Table 3.

First, LCC, SRCC and KTAU have coefficients close to 1 with each other in both utterance- and system-levels. However, MSE behaves differently from the above 3 metrics, as evidenced by the relatively lower correlation (~ 0.87). This analysis suggests that future studies may choose to report, for example, only the SRCC, while still keeping the MSE. It is also of interest to develop a general metric that takes both metrics into account.

5. Conclusions

We have presented a description of the first VoiceMOS Challenge and a summary of the results. We have observed the overwhelming effectiveness of SSL-based models for this task. We have also observed the difficulty of unseen synthesis systems in the OOD setting, and of unseen speakers and listeners in the main track. Generalizing to a different listening test context and predicting scores precisely using a small amount of labeled data remains challenging, as shown by the wide range of MSE values in the OOD track. We have also observed the importance of well-matched training and test distributions. Our future challenge will focus on the high-scoring region of the MOS range. Given the high quality of present-day speech synthesis systems, this task is very relevant to speech synthesis researchers, especially if MOS predictors are to be useful for objectively evaluating experimental synthesis systems.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number 21J20920, JST CREST Grant Number JPMJCR19A3, JPMJCR18A6 and JPMJCR20D3, and by MEXT KAKENHI grants 21K11951 and 21K19808. This work was also supported by MOST 107-2221-E-001-008-MY3. We would like to thank the organizers of the Blizzard Challenge and the Voice Conversion Challenge as well as the creators of ESPnet-TTS for making their audio samples available.

6. References

[1] ITUT Recommendation, "Vocabulary for performance and quality of service," *International Telecommunications Union—Radiocommunication (ITU-T), RITP: Geneva, Switzerland*, 2006.

[2] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," pp. 749–752, 2001.

[3] S. Möller, F. Hinterleitner, T. H. Falk, and T. Polzehl, "Comparison of Approaches for Instrumentally Predicting the Quality of Text-to-speech Systems," in *Proc. Interspeech*, 2010.

[4] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Towards perceptual quality modeling of synthesized audiobooks-Blizzard Challenge 2012," in *Blizzard Challenge Workshop*, 2012.

[5] ——, "Quality Prediction of Synthesized Speech based on Perceptual Quality Dimensions," *Speech Communication*, vol. 66, pp. 17–35, 2015.

[6] T. Yoshimura, G. E. Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda, "A Hierarchical Predictor of Synthetic Speech Naturalness Using Neural Networks," in *Interspeech*, 2016, pp. 342–346.

[7] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a Non-intrusive Assessor of Naturalness-of-speech," in *Proc. NIPS End-to-end Learning for Speech and Audio Processing Workshop*, 2016.

[8] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545.

[9] G. Mittag and S. Möller, "Deep Learning Based Assessment of Synthetic Speech Naturalness," in *Proc. Interspeech*, 2020, pp. 1748–1752.

[10] Y. Choi, Y. Jung, and H. Kim, "Deep MOS Predictor for Synthetic Speech using Cluster-Based Modeling," in *Proc. Interspeech*, 2020, pp. 1743–1747.

[11] J. Williams, J. Rownicka, P. Oplustil, and S. King, "Comparison of speech representations for automatic quality estimation in multi-speaker text-to-speech synthesis," *Speaker Odyssey*, 2020.

[12] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNET: MOS Prediction for Synthesized Speech with Mean-Bias Network," in *Proc. ICASSP*, 2021, pp. 391–395.

[13] W.-C. Tseng, C.-Y. Huang, W.-T. Kao, Y. Y. Lin, and H.-Y. Lee, "Utilizing Self-supervised Representations for MOS Prediction," in *Proc. Interspeech*, 2021, pp. 2781–2785.

[14] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. ICASSP*, 2022.

[15] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" in *Proc. SSW*, 2021, pp. 183–188.

[16] V. Karaiskos, S. King, R. A. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge Workshop*. Citeseer, 2008.

[17] A. W. Black, S. King, and K. Tokuda, "The Blizzard Challenge 2009," in *Proc. Blizzard Challenge*, 2009, pp. 1–24.

[18] S. King and V. Karaiskos, "The Blizzard Challenge 2010," 2010.

[19] ——, "The Blizzard Challenge 2011," 2011.

[20] ——, "The Blizzard Challenge 2013," 2013.

[21] ——, "The Blizzard Challenge 2016," 2016.

[22] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *Proc. Interspeech*, 2016, pp. 1632–1636.

[23] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods," in *Proc. Odyssey*, 2018, pp. 195–202.

[24] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice Conversion Challenge 2020 - Intra-lingual semi-parallel and cross-lingual voice conversion - ," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 80–98.

[25] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "ESPnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit," 2020.

[26] Z. Wu, Z. Xie, and S. King, "The Blizzard Challenge 2019," in *Proc. Blizzard Challenge Workshop*, vol. 2019, 2019.

[27] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep Learning-based Non-Intrusive Multi-Objective Speech Assessment Model with Cross-Domain Features," *arXiv preprint arXiv:2111.02363*, 2021.

[28] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "LDNet: Unified Listener Dependent Modeling in MOS Prediction for Synthetic Speech," in *Proc. ICASSP*, 2022.

[29] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS 2020*, 2020.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[31] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," in *Proc. ACL*, 2020, pp. 8342–8360.