Multi-Perspective Transfer Learning for Automatic MOS Prediction of Low Resource Language

Pengkai Yin, Rui Liu*, Feilong Bao, Guanglai Gao

College of Computer Science, Inner Mongolia University, Hohhot, China
liurui_imu@163.com

***Abstract*—Mean Opinion Score (MOS) prediction is the task to automatically evaluate synthesized speech by a neural network that emulates a human listening test. Traditional automatic MOS prediction typically focused on mainstream languages, such as English, due to large available data. However, for low-resource languages, there is no large-scale MOS prediction data, that hinders the study of those languages. In this paper, we propose a novel *Multi-Perspective Transfer Learning* (*MPTL*) training scheme with a new small-scale Mongolian MOS prediction dataset *MonMOS*. MPTL includes *Feature Transfer* and *Model Transfer* to transfer knowledge from the mainstream languages to low-resource language from different perspectives. The experimental results on the *MonMOS* show that the MPTL outperforms the standard direct training scheme with classical architecture. *We will release the pre-trained models and MonMOS dataset at: https://github.com/Ai-S2-Lab/MPTL-MOS* .**

***Index Terms*—MOS Prediction, Low-Resource Language, Transfer Learning**

I. Introduction

Mean Opinion Score (MOS) prediction [1] aims to assess the overall quality of the speech generated from Text-to-Speech (TTS) et al. with the help of neural networks [2], [3]. Note that objective evaluations, such as Mel-cepstral distance (MCD) [4], are not always correlated with human perception [5], [6], while manual MOS evaluations are often time-consuming and labor-intensive [1]. Therefore, automatic MOS prediction becomes an appealing alternative to subjective evaluation.

The neural network approach to automatic MOS prediction has made much progress. One influential work is MOSNet [1], which predicts MOS from spectrograms using a CNN and BLSTM-related architecture. Furthermore, MBNet [3] and LDNet [7] both explored explicitly learning the listener bias in the MOS data. Furthermore, Shen et al. [8] incorporated the Self-Supervised Learning (SSL) based representations [9] into MOS prediction and achieved remarkable performance. Despite the progress, the above works just focus on mainstream languages [10], [11] such as English, and their remarkable performance cannot be separated from the support of large-scale data, such as VCC2016 [12], VCC2018 [13], BVCC [2], BC2019 [14] and ASV2019 [15] et al. However, for low-resource languages, there is no large-scale MOS prediction data, and it is difficult to achieve encouraging results on small-scale data, which limits the development of TTS for such languages [11].

To address this issue, inspired by transfer learning [16], we propose a *Multi-Perspective Transfer Learning* (MPTL) training scheme for MOS prediction of low-resource language. MPTL includes *Feature Transfer* and *Model Transfer*. Specifically, 1) *Feature Transfer* aims to transfer knowledge from the speech self-supervised model of mainstream languages to low-resource language to learn robust acoustic feature representations for low-resource language; 2) *Model Transfer* seeks to transfer knowledge from the MOS prediction model of mainstream language to low-resource language to learn robust model starting points for low-resource language. Last but not least, we propose a new MOS prediction dataset *MonMOS* for the Mongolian language [17], a representative of low-resource language. The experimental results on *MonMOS* suggest that the proposed *MPTL* is able to efficiently transfer knowledge from mainstream language to low-resource language to achieve superior MOS prediction performance. The main contributions of this work can be summarized as follows:

- We propose a novel multi-perspective transfer learning scheme for automatic MOS prediction of low-resource language.To our knowledge, this is the first in-depth study of the training strategy for automatic MOS assessment of low-resource language;
- The MPTL consists of *Feature Transfer* and *Model Transfer* to perform robust SSL feature learning and accurate MOS prediction;
- The experimental results on MonMOS validated our MPTL.

In the rest of this paper, we first introduce the methodology of MPTL-MOS in Section II. Afterward, we present the experimental setup in Section III, which includes the dataset, the baseline, and the implementation details. We also show all the experimental results and conduct in-depth analyses in Section III. Finally, we conclude this paper and discuss future work in Section IV.

Rui Liu is corresponding author. The research by Rui Liu was funded by the Young Scientists Fund of the National Natural Science Foundation of China (No. 62206136), Guangdong Provincial Key Laboratory of Human Digital Twin (No. 2022B121201 0004), and the "Inner Mongolia Science and Technology Achievement Transfer and Transformation Demonstration Zone, University Collaborative Innovation Base, and University Entrepreneurship Training Base" Construction Project (Supercomputing Power Project) (No.21300-231510).

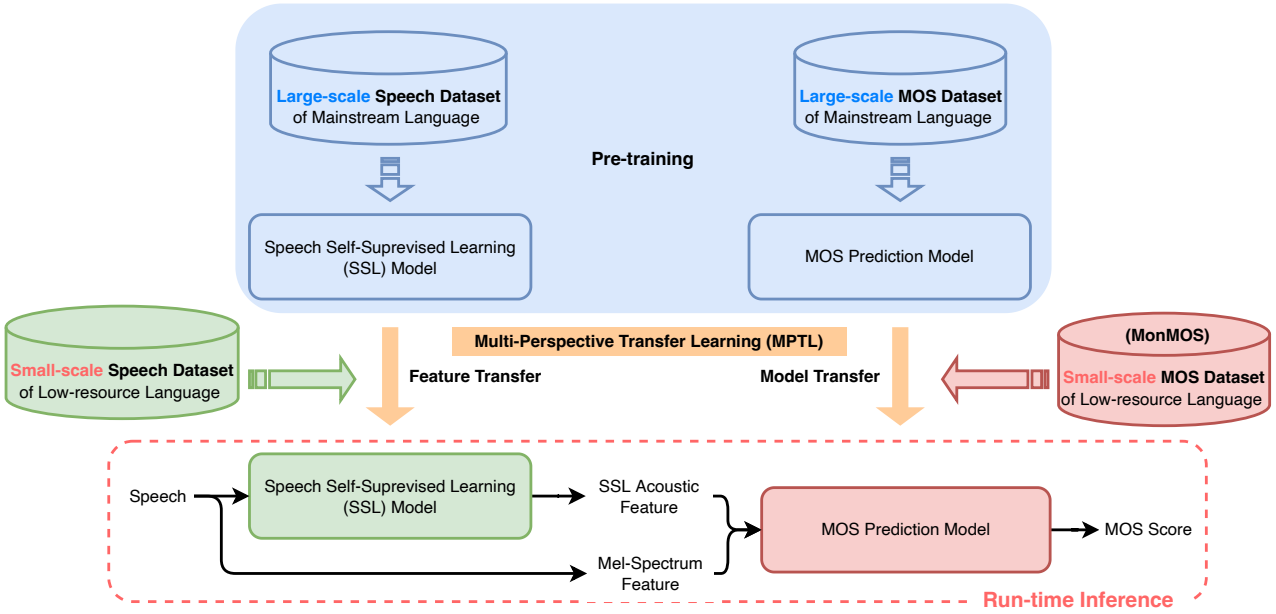

Fig. 1. The overall workflow of MPTL. Specifically, the backbone of MOS prediction network of low-resource language includes the *Speech Self-Supervised Learning Model* and *MOS Prediction Model*. In face of small-scale speech and MOS datasets of low-resource language, the *Feature Transfer* aims to transfer the knowledge from the speech SSL model of mainstream language, that pre-trained with a large-scale speech dataset, to that of low-resource language, while the *Model Transfer* aims to transfer the knowledge from the MOS prediction model of mainstream language, that pre-trained with a large-scale MOS dataset, to that of low-resource language.

II. MPTL: Methodology

We first describe the network backbone of our MPTL, then explain the MPTL workflow. The first perspective is *Feature Transfer*, that aims to transfer the knowledge from the speech SSL model of mainstream language, that pre-trained with a large-scale speech dataset, to that of low-resource language, to allow for extracting the robust SSL acoustic feature for low-resource language. The second perspective is *Model Transfer*, which seeks to transfer the knowledge from the MOS prediction model of mainstream language, that pre-trained with a large-scale MOS dataset, to that of low-resource language, thus building a robust mapping between the audio features and the MOS score for low-resource language.

A. Network Backbone

As shown in Fig.1, our MPTL adopts the *Speech SSL Model* and the *MOS Prediction Model* as the backbone. Note that the *Speech SSL Model* generates the robust SSL feature [18] by reading the input speech, while the *MOS Prediction Model* takes both the SSL feature and traditional Mel-spectrum feature to predict the MOS score.

For the *Speech SSL Model*, the key structure is a Masked Acoustic Model (MAM) [19] that achieves unsupervised speech representation learning. Given the masked frames, the MAM learns to reconstruct and predict the original frames. In this work, we select the Mockingjay model [20] to extract the SSL feature. Specifically, Mockingjay uses multi-layer transformer encoders and multi-head self-attention [21] to achieve bidirectional encoding, thus considering past and future contexts at the same time. During training, randomly select 15% of the input frames, and the Mockingjay model predicts the selected frames based on its left and right context. The L1 Loss is used to minimize the reconstruction error between prediction and ground-truth frames on the selected 15%. More details are referred to [20]. In a nutshell, assume the input speech is X, the Speech SSL Model reads X and outputs the SSL feature $\mathcal{H}$:

$$\mathcal{H} = \tilde{\Theta}_{ll}^{SSL}(X) \tag{1}$$

where $\tilde{\Theta}_{ll}^{SSL}$ means the model parameters of the speech SSL model.

For the *MOS Prediction Model*, we adopt the CNN, BiLSTM, and their mixup as three comparative structures as detailed in Section III, since these are the dominant structures in the field of MOS prediction [1]. Following [1], we combine the utterance-level Mean Square Error (MSE) [22] and frame-level MSE as the final loss function. As mentioned before, the mathematical expression for the MOS prediction is as follows:

$$Y = \tilde{\Theta}_{ll}^{MOS}(concat(\mathcal{H}, f)) \tag{2}$$

where $\tilde{\Theta}_{ll}^{MOS}$ means the model parameters of the MOS prediction model, f is the hand-crafted Mel-spectrum feature, Y indicated the final MOS score of low-resource language. The $concat$ function seeks to concatenate two features by feature dimension [8]. In run-time inference, the trained *Speech SSL Model* and *MOS Prediction Model* are combined together to predict the final MOS score for the input speech.

B. MPTL Workflow

As shown in Fig. 1, the parameters of each module, that are $\tilde{\Theta}_{SSL}$ and $\tilde{\Theta}_{MOS}$, in the network backbone are obtained by the proposed MPTL strategy separately. We first pre-train each module based on large-scale speech and MOS datasets in mainstream languages. After that, with the help of small-scale speech and MOS datasets of low-resource languages, we transfer the pre-trained knowledge to the corresponding modules of low-resource languages respectively.

1) Pre-training: In the pre-training stage, we first train the model parameters of *Speech SSL Model* and *MOS Prediction Model* for mainstream languages to provide a stable initial state for subsequent transfer learning. We denote the large-scale speech and MOS datasets of mainstream language as LSD_{ml} and LMD_{ml} respectively. To be consistent with the low-resource language, the *Speech SSL Model* and *MOS Prediction Model* share the same structure with the corresponding modules of low-resource language in Section II-A. In this way, we obtain the model parameters of *Speech SSL Model* and *MOS Prediction Model*, that are $\hat{\Theta}_{ml}^{SSL}$ and $\hat{\Theta}_{ml}^{MOS}$, for mainstream languages.

2) Feature Transfer: Feature transfer aims to transfer the knowledge from the speech SSL model of mainstream language, that pre-trained with a large-scale speech dataset, to that of low-resource language. Specifically, we denote the small-scale speech dataset of low-resource language as SSD_{ll}. Note that we use the trained $\hat{\Theta}_{ml}^{SSL}$ to initialize the Θ_{ll}^{SSL}, then fine-tune the Θ_{ll}^{SSL} with SSD_{ll} to conduct feature transfer. In this way, we obtain the trained $\tilde{\Theta}_{ll}^{SSL}$ for low-resource language, which allows extracting the robust SSL acoustic feature $\mathcal{H}$ for the low-resource language.

3) Model Transfer: Model transfer seeks to transfer the knowledge from the MOS prediction model of mainstream language, that pre-trained with a large-scale MOS dataset, to that of low-resource language. Specifically, we denote the large-scale MOS dataset of low-resource language as $SMD_{ll} = \{X_{ll}, Y_{ll}\}$, X_{ll} and Y_{ll} mean the audio files and their MOS score respectively. Note that we use the trained $\hat{\Theta}_{ml}^{MOS}$ to initialize the Θ_{ll}^{MOS}, then fine-tune the Θ_{ll}^{MOS} with SMD_{ll} to conduct feature transfer. In this way, we obtain the trained $\tilde{\Theta}_{ll}^{MOS}$ for low-resource language, which allows learning the mapping between the audio features and the MOS score.

With the help of MPTL, our MOS prediction network allows for making good use of the knowledge of large-scale data in mainstream language to achieve satisfactory performance for the task of MOS prediction in low-resource languages.

III. Experiments and Results

A. Datasets

To conduct pre-training and transfer learning in our approach, we incorporate various datasets. Specifically, we use English and Mongolian datasets to represent mainstream and low-resource languages, respectively. This helps us gain insights into effective transfer learning strategies for similar low-resource languages.

- LSD_{ml}: LibriSpeech [23] is treated as the large-scale speech dataset of mainstream language. we just use the train-clean-360 subset to conduct the SSL model pertaining.
- LMD_{ml}: VCC2018 [13] is treated as the large-scale MOS dataset of mainstream language that contains 20580 audios submitted by 38 different systems. A total of 267 expert judges are involved in VCC2018 and each audio is scored by 4 judges. We randomly split the dataset into training, validation and test sets with a size of 13580, 3000 and 4000.
- SSD_{ll} and SMD_{ll}: To facilitate MPTL, we present the MonMOS dataset [1] as a small-scale collection of speech and MOS datasets. Specifically, MonMOS is derived from the final submission results of the Low-Resource Mongolian Text-to-Speech Synthesis Challenge 2022 (NCMMSC2022-MTTSC) [2], comprising 2800 audio samples, about 4.04 hours, submitted by 13 teams. The evaluation process involved 20 expert judges who participated in the MOS listening test, with each audio sample being scored by 4 judges. To ensure a robust evaluation, we partitioned the dataset randomly into training, validation, and test sets, with proportions of 65%, 15%, and 20% respectively. By introducing the MonMOS dataset, we aim to facilitate research and advancements in MOS prediction of low-resource language.

B. Comparative Study

1) Various Training Schemes: To validate the proposed MPTL, We develop four training methods for a comparative study.

- 1) **Directly Training (DT)**: We follow MOSNet [1] and directly train the MOS prediction network with the MonMOS dataset without any pre-training and transfer learning;
- 2) **DT + Model Transfer**: We just conduct *Model Transfer* on the basis of DT, that means the $\tilde{\Theta}_{ll}^{MOS}$ is transferred by the $\hat{\Theta}_{ml}^{MOS}$;
- 3) **DT + Feature Transfer**: We just conduct *Feature Transfer* on the basis of DT, that means the $\tilde{\Theta}_{ll}^{SSL}$ is transferred by the $\hat{\Theta}_{ml}^{SSL}$ and the $\tilde{\Theta}_{ll}^{MOS}$ is trained from scratch. The input feature of MOS prediction model includes SSL feature and the Mel-spectrum feature;
- 4) **w/o *concat*** is an ablation study for feature concatenation in which we just use the SSL feature. Note that the first three baselines are treated as the ablation study for the *Feature Transfer* and *Model Transfer*.

2) Various Architecture: Following MOSNet [1] [3], we adopt various architectures, including CNN, BLSTM, and CNN+BLSTM as the main component of MOS prediction

[1]Dataset link: https://github.com/Ai-S2-Lab/MPTL-MOS

[2]http://mglip.com/challenge/NCMMSC2022-MTTSC/index.html

[3]Although this model was originally proposed in the field of voice conversion, it has also played a wide role in the field of TTS, and it is also suitable as the backbone.

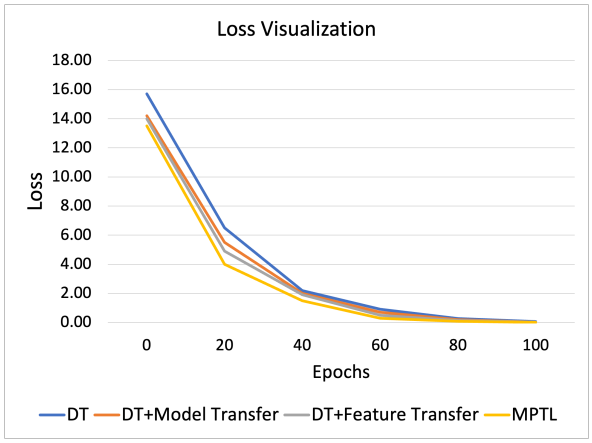

Fig. 2. Loss visualization of various models.

model. At last, we use two FC layers and a global averaging operation to obtain the utterance-level MOS score.

C. Experimental Setup

We extract 80-channel mel-spectrum features with a frame size of 50 ms and 12.5 ms frame shift. All speech samples are resampled to 16 kHz. For the speech SSL model, we just follow the configuration of Mockingjay [20]. For the MOS prediction model, the CNN module consists of 4 Conv2D blocks with filter size [16, 32, 64, 128], respectively. Each block includes 3 Conv2D layers with strides shape {[1,1], [1,1], [1,3]}, respectively. Each direction of BiLSTM contains 128 cells. The pre-training steps of the speech SSL model and MOS prediction model are 200000 and 100 epochs. We conduct feature transfer and model transfer with 100000 and 100 epochs respectively. All systems are trained by the Adam [24] optimizer with a learning rate of 0.0001. The dropout rate is set to 0.3. We set the batch size to 64. All DT systems are trained with 100 epochs. The MSE [22] value is treated as the objective metric.

D. Experimental Results

We report the MSE results of the comparative study as Table I. It has negatively oriented scores and lower values are better. We conducted a comprehensive analysis of the results in the following three areas.

1) Comparison of training scheme: As shown in the fifth row of Table I, the MPTL achieves the best performance over all baselines, no matter what architecture. For example, for CNN, the MSE of MPTL method is 0.09, while the other baselines achieve 0.17, 0.11, 0.13, and 0.31 respectively. In addition, MPTL gains the best MSE with 0.07 for BiLSTM and 0.09 for BiLSTM. Comparing MPTL and DT, the gap of 0.08 demonstrates the effectiveness of our proposed MPTL, which can achieve remarkable MOS prediction performance with small-scale dataset for low-resource language. Comparing MPTL with *DT + Model Transfer* and *DT + Feature Transfer* can be viewed as an ablation experiment for *Model Transfer* and *Feature Transfer*, we can clearly observe that both transfer

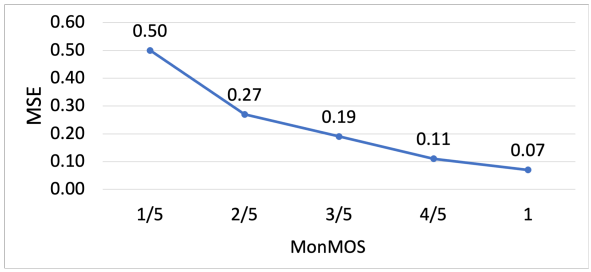

Fig. 3. MSE results of different data sizes of MonMOS.

learning skills result in performance improvements. The MSE value of the last row suggests that concatenating the SSL and Mel-spectrum features allows for robust acoustic representation by leveraging their complementary nature. We note that the last ablation baseline just uses SSL feature and performs the lowest MSE. Although similar phenomena have appeared in other work [25], we thought it might be useful to try some other SSL representations, such as HuBERT, etc., to explore this issue in the future.

2) Comparison of architecture: For architecture comparison, we find inconsistent findings with MOSNet [1]. The results in Table I show that BiLSTM, instead of CNN+BiLSTM, performs the best performance. This reason may be due to the fact that the data scale of MonMOS is so small that it does not match the relatively complex structure. BiLSTM has the ability to integrate long-term time dependencies and sequential characteristics into representative features, thus outperforming CNN.

3) Loss Visualization: We visualize the loss function of different models, as shown in Fig. 2. It can be observed that our method converges quickly and achieves lower loss values. The curve of loss variation indicates that our method improves the performance of MOS prediction, resulting in more accurate MOS prediction results. Therefore, our method is considered feasible for MOS prediction tasks in low-resource languages.

E. Discussion of Data Requirement

To further explore the relationship between our MPTL and data requirements, we aliquoted the MonMOS data into 5 parts (each part includes a duration of about 48 minutes) to compare the effect of MPTL at different data sizes. [1/5, 2/5, 3/5, 4/5, 1] represent data sizes of 560, 1020, 1920, 2240, and 2800 samples respectively. The durations for each data size are about 48, 96, 144, 192, and 240 minutes. We report the results in Fig. 3. We can observe that the MSE gradually decreases as the data size increases. We venture to guess that the model will keep getting better if the data size keeps increasing. However, increasing data size is not a smart way for us to solve the low-resource problem, In the future, we will continue to investigate how to achieve more efficient transfer learning methods with fewer data.

IV. Conclusion

This paper presents a novel Multi-Perspective Transfer Learning (MPTL) training scheme for automatic MOS predic-

TABLE I
MSE RESULTS OF THE COMPARATIVE STUDY. (·) INDICATES THE INPUT AUDIO FEATURE OF THE MOS PREDICTION MODEL IN THAT METHOD.

Method	**CNN**	**BiLTSM**	**CNN+BiLSTM**
DT (Mel) [1]	0.17	0.13	0.15
DT+Model Transfer (Mel)	0.11	0.10	0.11
DT+Feature Transfer (Mel+SSL)	0.13	0.09	0.12
MPTL (Mel+SSL)	**0.09**	**0.07**	**0.09**
w/o *concat* (SSL)	0.31	0.19	0.22

tion of low-resource language. MPTL successfully transfers acoustic knowledge learned from a large-scale dataset of mainstream language to the model of low-resource language by means of feature transfer and model transfer, achieving impressive performance. To contribute to the development of the field, we choose Mongolian, a representative of low-resource language, as the research object and release a small-scale Mongolian MOS prediction dataset, called MonMOS. Comprehensive experimental comparisons and analyses validated the effectiveness of our method. As per our knowledge, the proposed MPTL is the first study of MOS prediction training method for low-resource language. In future work, we intend to explore more SSL models for feature transfer and more network structures for model transfer, and further extend more languages to validate the effectiveness of the method.

REFERENCES

[1] 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," *Proc. Interspeech 2019*, pp. 1541–1545, 2019.

[2] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8442–8446.

[3] 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 *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 391–395.

[4] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1. IEEE, 1993, pp. 125–128.

[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: fast, robust and controllable text to speech," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 3171–3180.

[6] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2020.

[7] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "Ldnet: Unified listener dependent modeling in mos prediction for synthetic speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 896–900.

[8] K. Shen, D. Yan, and L. Dong, "Msqat: A multi-dimension non-intrusive speech quality assessment transformer utilizing self-supervised representations," *Applied Acoustics*, vol. 212, p. 109584, 2023.

[9] F. S. Oliveira, E. Casanova, A. C. Júnior, L. RS Gris, A. S. Soares, and A. R. Galvão Filho, "Evaluation of speech representations for mos prediction," in *International Conference on Text, Speech, and Dialogue*. Springer, 2023, pp. 270–282.

[10] T. Sellam, A. Bapna, J. Camp, D. Mackinnon, A. P. Parikh, and J. Riesa, "Squid: Measuring speech naturalness in many languages," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[11] P. Do, M. Coler, J. Dijkstra, and E. Klabbers, "Resource-efficient fine-tuning strategies for automatic mos prediction in text-to-speech for low-resource languages," *arXiv preprint arXiv:2305.19396*, 2023.

[12] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016." in *Interspeech*, 2016, pp. 1632–1636.

[13] 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 *The Speaker and Language Recognition Workshop (Odyssey 2018)*. ISCA, 2018, p. 195.

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

[15] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[16] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[17] R. Liu, B. Sisman, F. Bao, J. Yang, G. Gao, and H. Li, "Exploiting morphological and phonological features to improve prosodic phrasing for mongolian speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 274–285, 2020.

[18] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[19] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[20] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] O. Köksoy, "Multiresponse robust design: Mean square error (mse) criterion," *Applied Mathematics and Computation*, vol. 175, no. 2, pp. 1716–1729, 2006.

[23] 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.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] A. Kunikoshi, J. Kim, W. Jun, and K. Sjölander, "Comparison of speech representations for the mos prediction system," *arXiv preprint arXiv:2206.13817*, 2022.