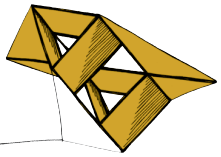

Effects of Large Multi-Speaker Models on the Quality of Neural Speech Synthesis

Lukáš Vladař[1]

1 Introduction

These days, speech synthesis is usually performed by neural models (Tan et al., 2021). A neural speech synthesizer is dependent on a large number of parameters, whose values must be acquired during the process of model training. In many situations, the result of training can be improved by *fine-tuning* a *pre-trained model*, i.e. using the parameter values of a model which has been trained using different training data to initialize the parameters of the target model before the training process begins (Zhang et al., 2023).

In the field of speech synthesis, a pre-trained model is a speech synthesizer which has been trained to synthesize the voice of another speaker. Furthermore, we can use a *multi-speaker* pre-trained model, which has been trained using speech recordings of multiple speakers, so it should contain general knowledge about human speech.

This paper describes how the number of speakers used to train a pre-trained model affects the quality of the final synthetic speech. We used a single-speaker model as well as two multi-speaker models for fine-tuning and we compared the obtained models in a listening test.

2 Available dataset and pre-trained models

To train the models, we used a dataset containing 90 minutes of speech of a Czech non-professional male speaker. For fine-tuning, three pre-trained models were available: a single-speaker model trained using recordings of a Czech professional speaker with a total duration of 16.7 hours and two Czech multi-speaker models. One of the multi-speaker models was trained using high-quality studio recordings of 6 professional speakers (3 male and 3 female) with a total duration of 100.9 hours, while the other one was trained using a large dataset containing recordings of 1,116 speakers totalling 267.7 hours. However, this large dataset was not primarily intended for speech synthesis, as it contained lower quality recordings recorded by non-professional speakers.

3 Model training

We used each of the three pre-trained models to train a speech synthesizer. For this experiment, VITS, which is an end-to-end TTS model proposed by Kim et al. (2021), was chosen. All the models were trained using batches of 32 recordings up to approximately 300,000 training steps. We used the AdamW optimizer and the learning rate was set to 0.0005.

[1] student of the doctoral degree program Cybernetics, e-mail: vladarl@kky.zcu.cz

4 Experiment evaluation

The speech synthesizers obtained by fine-tuning were compared in a MUSHRA listening test, which is described in the recommendation Method for the subjective assessment of intermediate quality level of coding systems (2014). The test was performed by eight listeners. It contained 20 utterances, with each utterance represented by four recordings (three synthesized ones and an original speaker's recording). The listeners were asked to rate each recording with a number from 0 to 100. The final rating of each model by a particular listener was obtained as the average of all scores given to the model by that listener throughout the whole test.

We found out that some listeners choose their score only from a small range of values, while others use the whole scale from 0 to 100. Therefore, we normalized the scores provided by each listener to match the mean and variance of the average listener's scores.

5 Results

The results of the listening test are shown in Figure 1. As expected, the highest MUSHRA score was reached by original recordings of the speaker. The second best result was obtained by the model which had been created by fine-tuning a large multi-speaker model. Conversely, fine-tuning a single-speaker model turned out to be the least effective training strategy.

The results of the experiment imply that the more speakers we use to train a pre-trained model, the better results we can get by fine-tuning that model. Thus, large multi-speaker models trained using recordings of thousands of speakers might be very beneficial for speech synthesis.

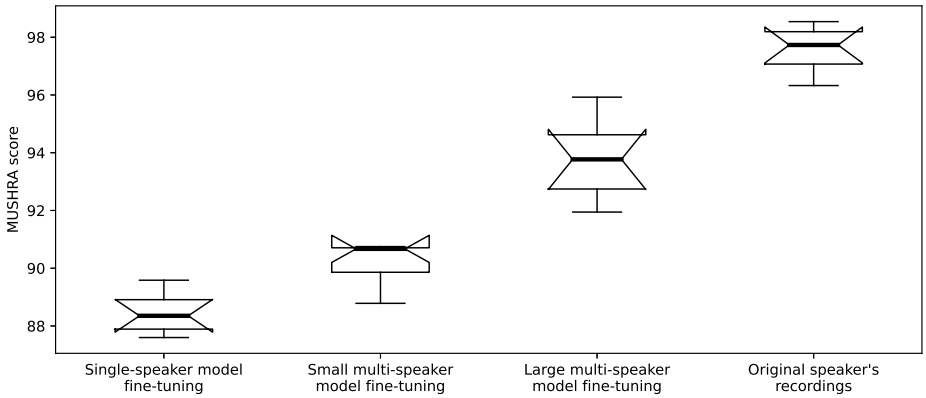

Figure 1: Results of the MUSHRA listening test visualized in a boxplot

References

International Telecommunication Union (2014). *Method for the subjective assessment of intermediate quality level of coding systems*. Available at `https://www.itu.int/rec/R-REC-BS.1534-3-201510-I/en`.

Kim, J., Kong, J., Son, J. (2021). Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *Proceedings of the 38th International Conference on Machine Learning*.

Tan, X., Qin, T., Soong, F., Liu, T.-Y. (2021). *A Survey on Neural Speech Synthesis*. Available at `https://arxiv.org/abs/2106.15561`.

Zhang, A., Lipton Z.C., Li, M., Smola A.L. (2023). *Dive into Deep Learning*, pp. 622–629. Available at `https://arxiv.org/abs/2106.11342`.