

## A DATA

In this section, we describe details of the data usage in training and evaluating VoiceTuner.

- For self-supervised pre-training, Librilight [18] contains 60K hours of unlabeled speech from audiobooks in English, and WenetSpeech [47] include 100K hours of speech in mandarin.
- For zero-shot text-to-speech, LibriTTS [46] dataset is included.
- For instruction text-to-speech, we use the dataset PromptSpeech [9].
- For singing voice synthesis, We use the female-singer OpenCPOP [42], multi-singer dataset OpenSinger [14], and M4Singer [48] as the singing voice data.

## B MODEL CONFIGURATIONS

We list the model hyper-parameters of VoiceTuner in Table 10.

### B.1 MIDI-to-F0 Converter

Singing voice synthesis (SVS) is a task that generates singing voices from the given music score and lyrics like human singers. Following [28, 48], the SVS system typically includes the MIDI-to-F0 converter to predict F0 explicitly. Though the SVS system can be further improved with the direct MIDI condition and implicit F0 prediction, this is beyond our focus.

### B.2 Unit-based Vocoder

The generator of the unit-based vocoder is built from a set of look-up tables (LUT) that embed the discrete representation, and a series of blocks composed of transposed convolution and a residual block with dilated layers. We train the enhanced vocoder with the weighted sum of the least-square adversarial loss, the feature matching loss, and the spectral regression loss on mel-spectrogram, where the training objective formulation and hyperparameters follow Kong et al. [21], Lee et al. [24].

For speech generation, we train the vocoder with only the discrete unit sequences as input. For singing voice generation, we further include F0-driven source excitation to stabilize long-continuous waveforms generation following [15, 28].

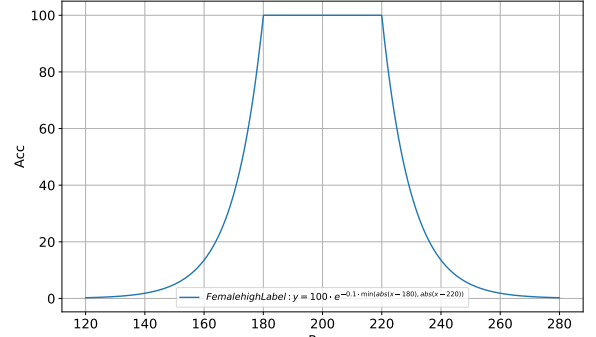
## C EVALUATION

### C.1 Objective Evaluation

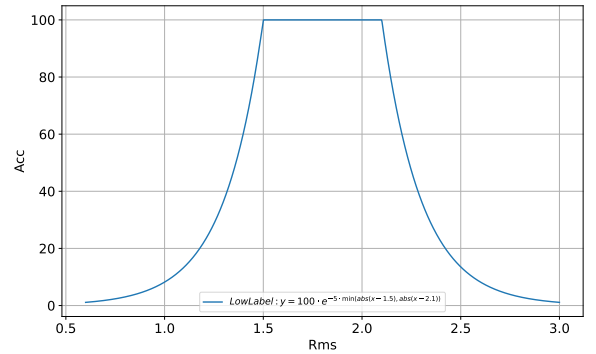
For controlling accuracies on volume, pitch, and speaking speed, considering that the values of generated singing may slightly deviate from the boundaries used for categorization, we adopt a soft-margin mechanism for accuracy calculation. Specifically, we take the accuracy of data falling within the correct range as 100, and calculate the accuracy with  $100 * \exp(-k\epsilon)$  for data outside the correct range, where  $\epsilon$  is the error between the data value and the boundary, and  $k$  is a hyper-parameter controlling the decay rate of accuracy at the margins, with larger  $k$  corresponding to faster decay. We take accuracy curves of high vocal-range of female, low speed, and medium volume as examples and illustrate them in Figure 5.

### C.2 Subjective Evaluation

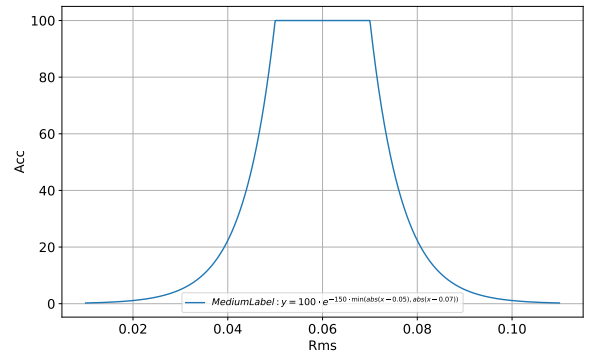
For audio quality evaluation, we conduct the MOS (mean opinion score) tests and explicitly instruct the raters to “(focus on examining the audio quality and naturalness, and ignore the differences



(a) Pitch.



(b) Speed.



(c) Volume.

Figure 5: Soft-margin accuracy curve.

of style (timbre, emotion, and prosody)). The testers present and rate the samples, and each tester is asked to evaluate the subjective naturalness on a 1-5 Likert scale.

For style similarity evaluation, we explicitly instruct the raters to “(focus on the similarity of the style (timbre, emotion, and prosody)

Hyperparameter		VoiceTuner
<b>VoiceTuner: Transformer</b>		
Global Base	Transformer Layer	16
	Transformer Embed Dim	768
	Transformer Attention Headers	12
	Number of Parameters	114 M
Global Medium	Transformer Layer	20
	Transformer Embed Dim	1152
	Transformer Attention Headers	16
	Number of Parameters	320 M
Global Large	Transformer Layer	24
	Transformer Embed Dim	1536
	Transformer Attention Headers	32
	Number of Parameters	830 M
Local	Transformer Layer	6
	Transformer Embed Dim	Same as global
	Transformer Attention Headers	8
	Number of Parameters	46/101/303 M
<b>BigVGAN Vocoder</b>		
BigVGAN Vocoder	Upsample Rates	[5, 4, 2, 2, 2, 2]
	Hop Size	320
	Upsample Kernel Sizes	[9, 8, 4, 4, 4, 4]
	Number of Parameters	121.6M

Table 10: Hyperparameters of VoiceTuner.

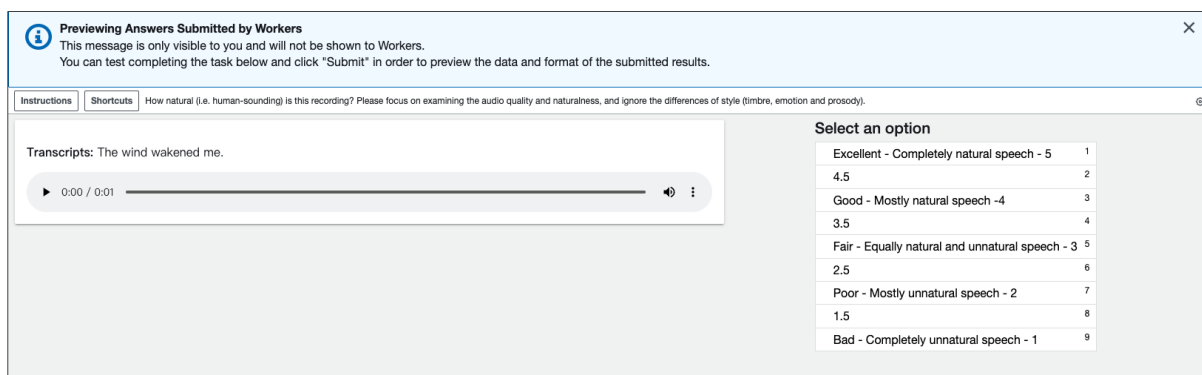
to the reference, and ignore the differences of content, grammar, or audio quality.”). In the SMOS (similarity mean opinion score) tests, we paired each synthesized utterance with a ground truth utterance to evaluate how well the synthesized speech matches that of the target speaker. Each pair is rated by one rater.

Our subjective evaluation tests are crowd-sourced and conducted by 20 native speakers via Amazon Mechanical Turk. The screenshots of instructions for testers have been shown in Figure 6. We

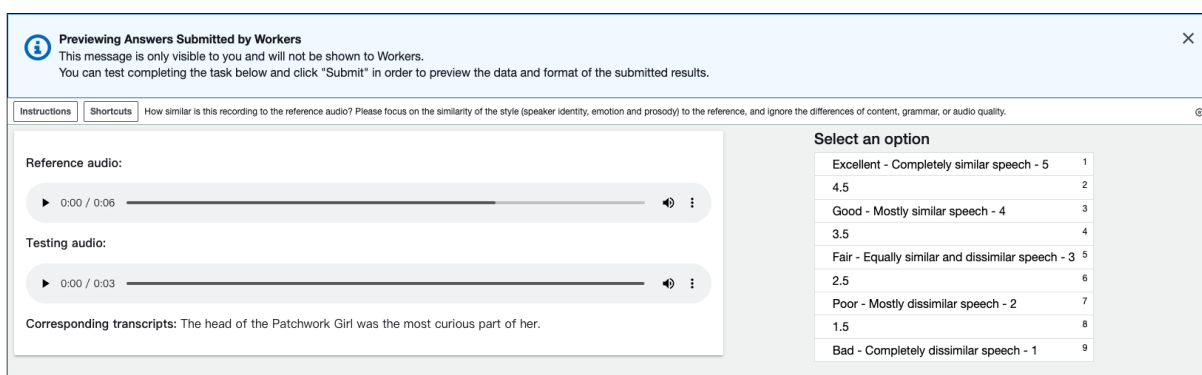
paid \$8 to participants hourly and totally spent about \$600 on participant compensation. A small subset of speech samples used in the test is available at <https://VoiceTuner.github.io/>.

D REPRODUCIBILITY STATEMENT

We will release our code in the future. The VoiceTuner model that we build upon is publicly available through the fairseq code repository [30]. To aid reproducibility, we have included a schematic overview of hyperparameters in Table 10.



(a) Screenshot of MOS testing.



(b) Screenshot of SMOS testing.

Figure 6: Screenshots of subjective evaluations.