Supplementary Materials: ExpressiveSinger: Multilingual and Multi-Style Score-based Singing Voice Synthesis with Expressive Performance Control

Anonymous Authors

1 DETAILS OF EVALUATIONS

Table 1 is a supplement to Section 5.2.2. It shows the subjective evaluation on synthesized quality of style, language, techniques in our system, compared with human singing. We ask users to rate according to both singing musicality, naturalness, audio quality, and pronunciation. In particular, when evaluating style and technique demo, we ask users to consider the realization of such style or technique input. We also ask native speaker users to evaluate the language naturalness.

Table 1: Subjective evaluation on quality of generated style, language, techniques, compared with human singing. CMOS scores for human singing are always 0.000, so we only show the CMOS scores of our system.

Style	CMOS
Pop	-0.322
Children	-0.407
Western opera	-1.021
Traditional Chinese Folk	-0.191
Jazz	-0.775
Teresa	-0.189
Language	CMOS
Chinese	-0.210
English	-0.461
Korean	-0.479
Italian	-0.826
Technique	CMOS
Lip trill	-0.269
Trill	-0.533
Vibrato	-0.207
Trillo	-1.016
Breathy	-1.328

Table 2 shows the results of subjective evaluation on zero-shot scenarios where the training dataset's singers had not previously attempted particular style and language (Section 5.2.4). "Seen language and style" means that the synthesized singer has previously sang the same language and style in the training data for different songs. Meanwhile, "zero-shot" means the synthesized singer has not ever sung such style or language before. "with Singer ID" means changing the singer embedding in acoustic model to singer ID, otherwise still using singer embedding. "concatenated phoneme set" means using the directly concatenated phoneme sets from all datasets instead of the combined phoneme set.

Table 2: Subjective evaluation on zero-shot scenarios where the training dataset's singers had not previously attempted particular style and language.

Scenarios	MOS
Seen language & style	3.635 ± 0.098
Zero-shot	3.029 ± 0.519
Zero-shot (with Singer ID)	2.826 ± 0.715
Zero-shot (concatenated phoneme set)	2.709 ± 0.634

2 DETAILS OF MODEL DESIGN

2.1 Detailed design of input condition context

Table 3 shows the detailed input condition context design for each expressive performance control model, as well as the acoustic model. It also describes the embedding layer architecture for different condition context. It is supplementary to Figure 3 in the paper.

2.2 First-stage rule-based algorithm in expressive timing model in Section 4.2.1

It splits the score-word timings counted in beats, into each phoneme's timing in seconds, without changing word boundary timings. It primarily accounts for the differences between vowel and consonant phonemes.

This algorithm addresses the mismatch between words and notes. First, when multiple words correspond to a single score note, we concatenate all word phonemes in sequence, which may include a rest note. Second, when one word matches multiple score notes, we assess the presence of multiple syllables; if the syllable count exactly matches the number of score notes, each syllable is sequentially assigned to a corresponding note. If the syllable count does not align with the note count, we merge the onsets and durations of all score notes into a single onset and duration. Following the above two steps, only one scenario remains: a note correspond to a list of phonemes.

Next, we distribute a note's duration among its phonemes, which include consonants, vowels, silences (SP), and probably breath sounds (AP) in some specific datasets. We assign durations of 0.03 seconds to each SP and AP, and 0.1 seconds to each consonant. The remaining duration is evenly allocated among the vowels. If this allocation results in vowel durations shorter than those of consonants, indicating insufficient note duration, we adjust the distribution: SPs are set to 0 seconds, APs to 0.02 seconds, and the remaining length is split with 40% to consonants and 60% to vowels.

Table 3: Detailed input condition context design. “Dimensionality” means the number of output dimensions for the embedding. “Input Representation” means representation for each data point in the input sequence. “Word boundary” here means whether current phoneme is the end of a word (0 or 1). “Phrase position” is current data point’s frame position within current phrase segment (counting from 1 to total frame length of the phrase. “Phoneme position” is the frame position within current phoneme. “Amplitude” and “F0” are the output amplitude envelopes and F0 curves from expressive performance model. “Quantized F0” is where F0 values in Hz are quantized into 256 discrete bins. “Score dur beat” means note duration counted in beats in the score. “Score onset sec” is the time onset counted in seconds for each score note. “Score position” represents the number of notes before current phoneme. Detailed definition and representation explanation please refer to Section 3.2 Data Representation.

Condition Context	Input Representation	Embedding Design	Dimensionality	Model Used
Singer embedding	256-dim float vector	None	256	Acoustic model
Singer ID	Integer	One-hot embedding	50	Timing F0 curves Amplitude envelopes
Language ID	Integer	One-hot embedding	4	Timing F0 curves Amplitude envelopes Acoustic model
Phoneme ID	Integer	Transformer	80	Timing F0 curves Amplitude envelopes Acoustic model
Dataset ID	Integer	One-hot embedding	5	Timing F0 curves Amplitude envelopes Acoustic model
Style genre ID	6-dim binary vector	None	6	Timing F0 curves Amplitude envelopes Acoustic model
Technique ID	Integer	One-hot embedding	17	Timing F0 curves Amplitude envelopes Acoustic model
Word boundary	0 or 1	None	1	Timing F0 curves Amplitude envelopes Acoustic model
Phrase position	Integer	Conv1x1	8	F0 curves Amplitude envelopes Acoustic model
Phoneme position	Integer	Conv1x1	4	F0 curves Amplitude envelopes Acoustic model
Amplitude	Float	Conv1x1	32	Acoustic model
F0	Float	Conv1x1	32	Acoustic model
Quantized F0	Integer	Fully-connected	32	Acoustic model
Score pitch	Integer	Fully-connected	32	Timing F0 curves Amplitude envelopes
Score dur beat	Integer	Fully-connected	8	Timing
Score onset sec	Float	Conv1x1	64	Timing
Score position	Integer	Conv1x1	16	Timing