Mandarin Singing Voice Synthesis with a Phonology-based Duration Model

Fu-Rong Yang*, Yin-Ping Cho*, Yi-Hsuan Yang†, Da-Yi Wu†, Shan-Hung Wu*, and Yi-Wen Liu*
* National Tsing Hua University, Hsinchu, Taiwan
E-mail: fjbcrs34@gmail.com, yinping.cho@outlook.com, ywliu@ee.nthu.edu.tw
Tel/Fax: +886-3-5162205
† Academia Sinica, Taipei, Taiwan
E-mail: yang@citi.sinica.edu.tw, ericwudayi2@gmail.com

***Abstract*—Singing voice synthesis (SVS) systems are built to generate human-like voice signals from lyrics and the corresponding musical scores. In most SVS systems, a neural network-based auxiliary duration model is employed to control the duration of phonemes. In this paper, a rule-based algorithm inspired by Mandarin phonology is proposed for the duration modeling in Mandarin SVS. Specifically, the algorithm infers the duration of an "initial" consonant by looking up syllables in an existing training set that begin with the same consonant and have similar note lengths, and then computing the average consonant duration. Around this, we employ a combination of Tacotron2 and Parallel WaveGAN as the backbone of our SVS system for their favorable data efficiency on small datasets. Experimental results show that the singing voice synthesized by the proposed duration model is more expressive than that of a learning-based model. Moreover, since Mandarin is a tonal language, the inclusion of tonality consideration further enhances the naturalness of the generated voices.**

I. Introduction

Over the recent years, machine learning and neural network (NN) models have become increasingly capable of generating human-like singing voices from musical scores with lyrics, a task we refer to as *singing voice synthesis* (SVS) in this paper. While traditional paradigms of performing SVS were based on concatenation techniques [1] and hidden Markov models (HMMs) [2], a variety of deep learning strategies have been adopted nowadays for SVS to improve the quality of the generated voices. For example, deep neural networks (DNNs) have been introduced to SVS [3] to learn the mapping from musical scores to acoustic features. Convolutional neural networks (CNNs) [4] and recurrent neural networks with long-short term memory (LSTM) cells [5] exhibited capabilities to capture the long-term dependencies of singing voices. Generative adversarial network (GAN) has also been adopted to alleviate the so-called over-smoothing problem [6].

In addition, taking advantage of the similarity between SVS and text-to-speech (TTS), researchers have also employed attention-based sequence-to-sequence (seq2seq) architectures, which have led to state-of-the-art TTS models [7]–[9], to SVS [10], [11]. As far as waveform generation is concerned, neural vocoders such as WaveNet [12], WaveRNN [13], and Parallel WaveGAN [14] have been used in SVS systems for converting acoustic features to time-domain audio samples, for producing natural-sounding singing voices [4], [10], [15], [16].

While lots of efforts have been made to improve the spectral fidelity of the synthesized singing voices, the impact of duration and time-lagging has not been thoroughly explored. Unlike the case in TTS, where the alignment error between the phoneme-level input to the frame-level output may not be harmful, such alignment error would be an important issue for SVS as it negatively affects the tempo and rhythms of singing. Coarse-grained alignment may even lead to skipped or repeated utterances that do not match the musical notes. Therefore, an auxiliary duration model is necessary for SVS. If we can expand beforehand the phoneme-level input sequence to a frame-level sequence with the same length as the target output sequence, the alignment between the encoder and decoder would be much easier [17], [18]. This has been demonstrated by Blaauw and Bonada [19], who showed that fine-grained alignment between the input sequences and the corresponding output acoustic features can be assured by informing the phoneme duration to a seq2seq-based SVS system.

In SVS, the duration model is commonly constructed by adopting a neural-network architecture without resorting to domain knowledge in linguistics. For example, a duration model may consist of several layers of LSTM [5], [10], [20] or CNN [21], followed by a post-processing step to constrain the voice timing associated with the note length. Although such networks are versatile for many purposes, they also rely heavily on the availability of a large annotated database. When the dataset is not sufficiently large, such duration models may not generalize well to unseen input scores.

Focusing on Mandarin SVS, here we propose a rule-based duration model relying on the phonology of Mandarin. Phoneme duration analyses were conducted within a recently annotated Mandarin singing dataset MPop600 [22] to provide the foundation of the rule-based algorithm. To evaluate its performance, the combination of Tacotron2 and Parallel WaveGAN is selected as the backbone of our SVS system due to their favorable data efficiency on small datasets. To compare the performance of different duration models, subjective and objective evaluations are conducted, and some audio samples for the listening test are available via this link[1]. Furthermore, tonality consideration and data augmentation are applied to

[1] https://furongyang.github.io/audio-demo-apsipa/

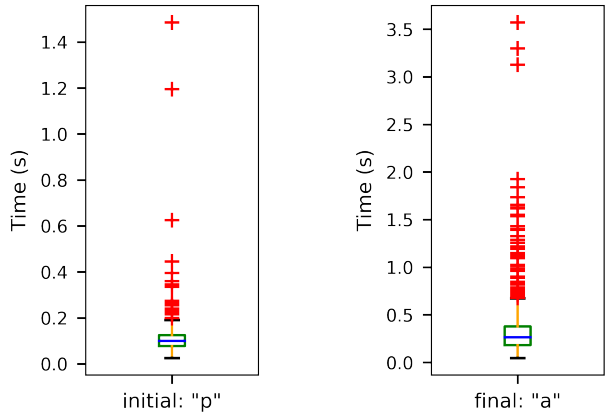

Fig. 1. Comparison on duration of *initials* and *finals* in singing. Through the quartiles, box plots visually show the distribution and skewness of phoneme durations in stretched Mandarin syllables.

our duration algorithm to see if robustness can be improved.

The paper is organized as follows. Section II introduces related background knowledge. Section III presents statistical analyses of phonology and the duration modeling algorithm. Section IV describes the design of our SVS system. In Section V, the experimental setup and the result evaluations are reported, and conclusions are given in Section VI.

II. Mandarin Phonology in a Nutshell

In most Mandarin SVS system, lyrics in a musical score are transcribed automatically by a linguistic processor from the character level to phonetic level. Specifically, each Chinese character represents a syllable, which can be decomposed into an *initial* followed by a *final*. While an *initial* can always be regarded as a consonant, a *final* can be a combination of a medial, a nucleus vowel, and a coda [23]. Therefore, strictly speaking, *initials* and *finals* are not equivalent to phonemes. For convenience, we casually refer to *initials* and *finals* as *phonetic components* (PhCs) in this paper. The list of PhCs in Mandarin Pinyin system can be looked up from many online resources[2]. Note that, in addition to the *initials* and *finals*, each Chinese character is also associated with a tone.

In Mandarin singing, when a character in lyrics is sung with a long-duration note, different regions of it are not stretched uniformly [24]. Linearly stretching on utterances may be unnatural since the vowel part (in the *finals*) should be stretched a lot more than the consonants in the *initials*. As a result, time-scale modification algorithms have been developed to achieve natural speech stretching [25], [26].

III. Duration Analysis

In SVS, a duration model is meant to predict the duration of each phoneme in the lyrics, since this information is not explicitly given by the musical scores. Without an auxiliary duration model, the number of final output frames is predicted directly from the alignment between phoneme-level and frame-level sequences, which is not precise enough. Once "how long a phoneme should sustain" is clearly informed, the duration error of SVS can be lower, improving the naturalness in the rhythm of singing. We provide an empirical evidence of this in Fig. 9, the result of subjective preference test in Section V.

In this paper, we attempt to determine the length of PhCs for SVS by an exemplar-based approach with simple phonological rules. In what follows, we present an analysis of stretching first, and then details of the way we develop our algorithm.

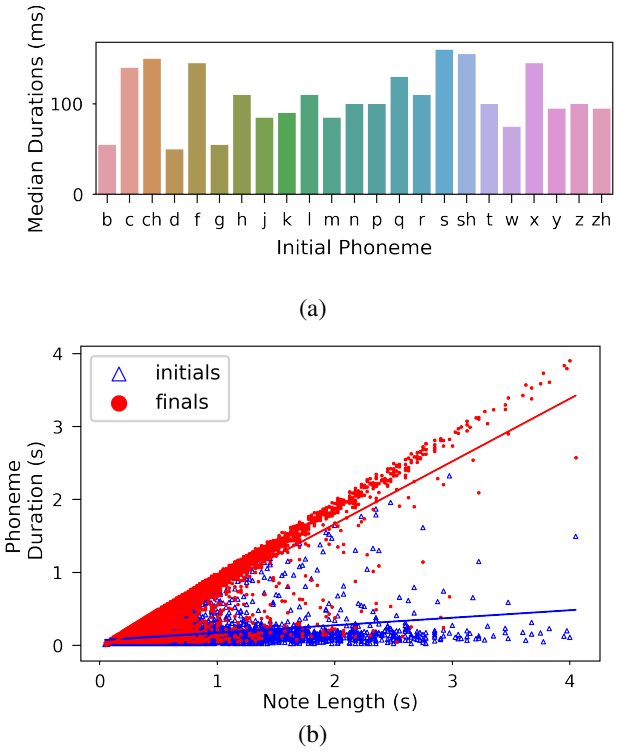

Fig. 2. Duration of phonetic components. (a) Median duration across all possible Mandarin *initials*, transcribed in Pinyin. (b) Duration vs. note length.

A. Phoneme Stretching Analysis

In human singing, when words are stretched by long-duration notes, different regions of a syllable are not stretched by the same ratio. For instance, Duan *et al.* [27] found that vowel sounds are stretched to maintain musical notes in singing based on their analysis of the NUS-48E corpus, an English singing and speech database. Here, we similarly analyzed the phoneme duration stretching in a Mandarin Chinese singing database MPop600 [22], which has duration annotations at the phoneme level. For instance, Fig. 1 displays the distribution of duration of the /p/ *initial* and the /a/ *final* in Mpop600. The box plots show that /a/ usually lasts longer than /p/, and the /p/ duration concentrates in a small range ($\leq$ 0.2 sec). Also, outliers are fewer in /p/ than in /a/, which corroborates our observation that *initials* would not stretch

[2]https://www.yellowbridge.com/chinese/pinyin-rules.php

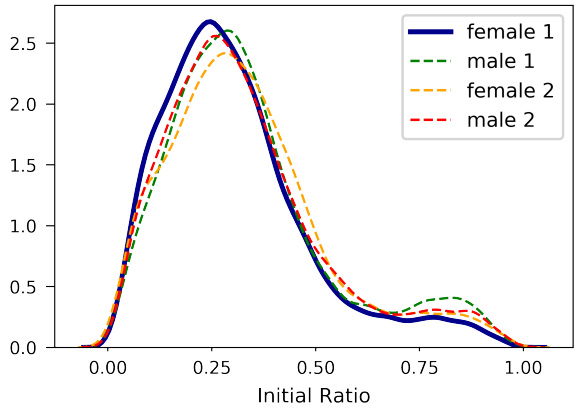

Fig. 3. Probability density function of all *initial ratios* from 4 singers.

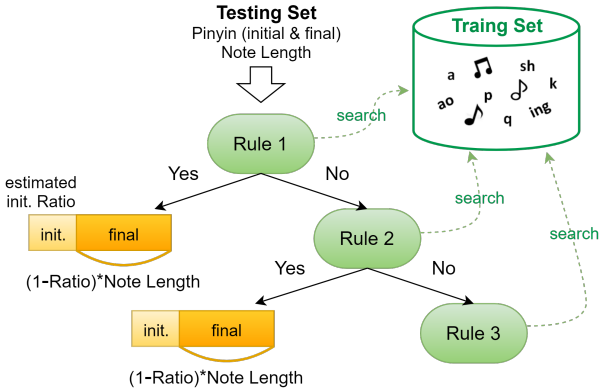

Fig. 4. Rule-based duration model illustrated by the decision tree.

much when a Chinese character is sung with a long note. The trend concluded from the /p/ and /a/ example displayed in Fig. 1 was consistently observed in most of the other *initials* and *finals* in MPop600, only with slight variations in their particular distributions.

With a thorough inspection of Fig. 2 (a), we see that the Mpop600 dataset covers all the *initials* relevant in modern Mandarin. More importantly, the figure confirms that different *initials* have significantly different durations, displaying the complexity of the duration prediction task. However, the duration of a phoneme is not solely decided by its own identity but mainly by the utterance length of the entire syllable. To investigate how *initials* and *finals* are stretched with various note lengths, we created a scatter graph for all the PhCs (see Fig. 2 (b)) sung by one female singer. Each dot represents one PhC duration with the corresponding note length, and the straight lines show the results of least-square regressions.

The slope of the regression line is 0.14 for the *initials*, and 0.86 for the *finals*. Along with the note length, the duration of *initials* (blue dots) increases and tends to maintain at a certain range. Also, as expected, the average duration of *finals* (red dots) keeps growing when the note length increases because vowel sounds are the dominant constituent of syllables. Combined, they show that the lengths of *initials* and *finals* are indeed highly correlated to note lengths.

Across different singers, we further found that the ratio of *initial* duration to note length (referred to as *initial ratio* hereafter) has little variation. This is shown in Fig. 3—the probability density functions for the *initial ratios* of the four singers indeed look similar, with a peak near 0.25.

B. Rule-based Algorithm

Our rule-based duration model treats the training set as a searching pool, which includes all the Mandarin PhCs, and the *initial ratio* is the major target of prediction. In MPop600, the phoneme durations were obtained by forced alignment[3] and are regarded as the ground truths in this research. As for unseen data, only musical scores were given and they come without the information of phoneme duration. We propose to determine PhC durations for unseen input data based on three rules; Fig. 4 illustrates the strategy with a decision tree, and Algorithm 1 demonstrates the conditions and implementations of these rules in detail. In Fig. 4, the input is a combination of $\{$*initial*, *final*$\}$ for a character and the corresponding note length. If the Pinyin transcription of a character contains only a *final*, a zero-*initial* token is assigned to it to ensure that all characters can be decomposed into two PhCs. To avoid confusion, some mathematical symbols in Algorithm 1 are first explained as follows,

- D and D': training set and testing set, respectively.
- p_i and q_i: *initial* and *final* decomposed from the i-th character in D.
- p'_i and q'_i: *initial* and *final* decomposed from the i-th character in D'.
- L_i: note length of the i-th character from D.
- L'_i: note length of the i-th character from D'.
- R_i: the predicted *initial ratio* of the i-th character.
- d_i: *initial* duration of the i-th character from D.
- d_i^{init} and d_i^{fin}: the predicted *initial* duration and *final* duration for the i-th character in D', respectively.

The following three rules are established to search the proper *initial ratio* from the training data.

1) **Rule 1***:* First, given $\{p'_i, q'_i, L'_i\}$ from one character in D', find out all entries in D with the same *initial*, same *final*, and the note length that does not differ from L'_i by more than 5 ms. Secondly, extract $\{d_i, L_i\}$ from all these entries in D and calculate the average *initial ratio*. The predicted *initial* duration is then obtained by multiplying this ratio with L'_i, while the *final* duration fills the remaining length.

2) **Rule 2***:* If no entry could be found in D to satisfy the requirement in rule 1, find out all entries with the same *initial* and the similar note length from the training data. We assume that with different *finals*, the *initial ratio* could still be effectively predicted. If the input meets the requirement of

[3]https://github.com/open-speech/speech-aligner

Algorithm 1 Rule-based duration model.

Input: Phonetic components (*initial* p'_i & *final* q'_i) of one character and and its note length L'_i in testing set D'.

Output: Durations d_i^{init} and d_i^{fin};

for all $(p'_i, q'_i, L'_i) \in D'$ **do**
  if Rule 1: $\mathcal{S} = \{(p_j, q_j, L_j) \in D : p_j = p'_i, q_j = q'_i, |L_j - L'_i| \leq 5 \text{ ms}\} \neq \emptyset$ **then**
    $R_i = \frac{1}{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{S}|} d_j / L_j$ ▷ Ratio of *initial* p'_i
    $d_i^{\text{init}} = R_i \times L'_i, \quad d_i^{\text{fin}} = (1 - R_i) \times L'_i$
  else if Rule 2: $\mathcal{S} = \{(p_j, q_j, L_j) \in D : p_j = p'_i, |L_j - L'_i| \leq 5 \text{ ms}\} \neq \emptyset$ **then**
    do step 3 $\sim$ 4
  else
    Rule 3: let $\mathcal{S} = \{(p_j, q_j, L_j) \in D : p_j = p'_i\}$
    for $k = 10, 20, ..., 50$ **do**
      find k-nearest neighbors of L'_i to form $S_k \subseteq \mathcal{S}$
      Calculate σ_k, the STD of d_j / L_j in S_k
    $k^\star = \arg\min_k \sigma_k$
    let $\mathcal{S} = S_{k^\star}$ and **do** step 3 $\sim$ 4
  return d_i^{init} and d_i^{fin};

rule 2, repeat the multiplication in rule 1 and return the output d_i^{init} and d_i^{fin}.

3) **Rule 3***:* If, still, no entries in D are found to satisfy the requirement in rule 2 due to the note length constraint (within 5 ms), we collect all entries of the same *initial*. From this set, find the k-nearest neighbors (KNN) whose note lengths are closest to L'_i and denote the subset as S_k. How many neighbors should be engaged is determined by the standard deviation (STD) of *initial ratio* in the subset S_k. Empirically, we set the value of k to be $10, 20, 30, 40$, or 50, and select $k = k^\star$ that produces the smallest STD. Here, we mention that the MPop600 covers all the Mandarin PhCs so rule 3 would always find a non-empty set.

Empirically, the actual percentage of terminating the search by rule 1, rule 2, and rule 3 is 88.0%, 11.2%, and 0.78%, respectively. Most of the times the search can be completed by rule 1 and 2 since the common note lengths are well balanced in the training data.

IV. System Design

As shown in Fig. 5, our proposed system consists of three modules: 1) A length regulator with a duration model, which expands the phoneme-level sequence to frame-level according to the phoneme duration; 2) A Tacotron2-based network which predicts a sequence of acoustic frames from an expanded frame-level input sequences; 3) A Parallel WaveGAN (PWG) which generates time-domain waveform samples conditioned on the mel-spectrogram.

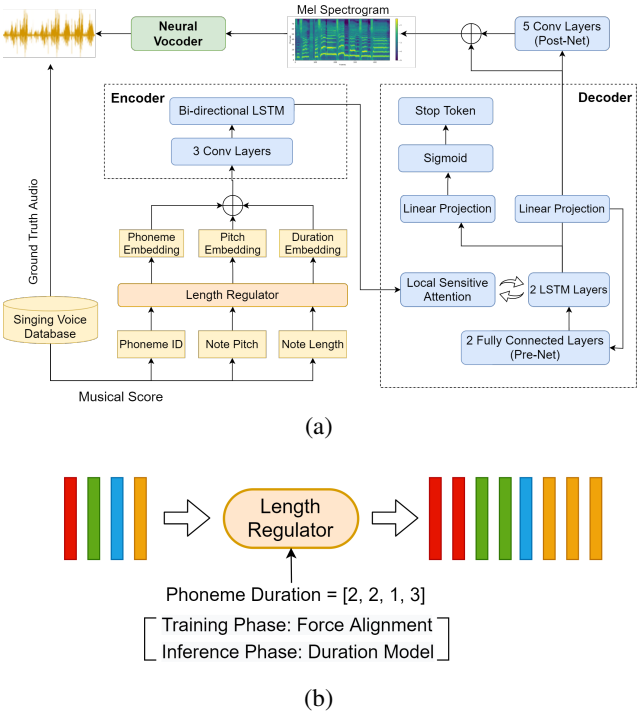

Fig. 5. System design. (a) The overview of the proposed Mandarin singing voice system. (b) The mechanism of how the length regulator works.

A. Length Regulator and Input Representation

The score information which includes the phoneme identity, the note pitch, and the note length is first fed into a length regulator module. This module, inspired by FastSpeech [17], utilizes the given phoneme durations to expand the length of the input matrices to the estimated number of final output frames by repetition. It should be clarified that the phoneme durations are given by forced alignment in the training phase, and obtained by the duration model in the inference phase (see Fig. 5(b)). This process mitigates the difficulty for the attention module by creating a quasi-one-to-one alignment between the ensuing hidden states and the predicted mel-spectrogram frames. Subsequently, the three expanded matrices are embedded separately in the same dimensional space and then summed together as the input sequence.

B. Acoustic Model

Our acoustic model follows the paradigm of mel-spectrogram prediction set out by Tacotron2 [8]. The encoder consumes the input sequence through a 3-layer convolutional network to embed the temporal context and integrate the information across the summed phoneme, pitch, and note length embeddings. A stack of 2 bi-directional LSTM layers then encodes the embedded sequence into sequential hidden states. For the decoder, 2 LSTM layers stack on a location-sensitive attention module [28], and their output is linearly projected

TABLE I
OBJECTIVE EVALUATION RESULTS OF DIFFERENT SYSTEMS

Model	Dur *er.*	Dur *cor.*	F0 *er.*	F0 *cor.*	MCD
model 1	38.31	0.40	6.83	0.73	15.08
model 2	10.37	0.95	6.27	0.88	8.26
model 3	8.63	0.96	5.02	0.90	7.12
model 4	8.74	0.96	4.68	0.95	7.03
model 5	**8.40**	**0.97**	**2.20**	**0.97**	**6.86**

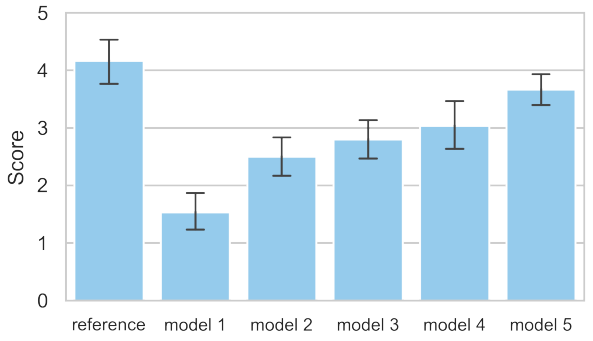

Fig. 6. Results of MOS test for each model and the reference. The error bars represent the 95% confidence intervals.

to be the predicted mel-frames. In this autoregressive setup, every predicted mel-frame passes through a 2-layer bottleneck module before feeding back to the decoder module to provide contextual information for predicting the next frame. Lastly, a convolutional post-net predicts the residual and adds it onto the predicted mel-spectrogram to get the refined final output.

C. Audio Synthesizer

The predicted mel-spectrogram is converted to waveform by a Parallel WaveGAN (PWG) neural vocoder trained on our dataset. PWG's modified non-autoregressive WaveNet generator makes it the state-of-the-art choice regarding audio quality and inference speed. Furthermore, its disjoint-training adversarial setup achieves high data efficiency, which is instrumental considering the scarcity of singing data. Specific to our SVS task, PWG has also shown superior robustness in synthesizing extended notes compared to its counterparts MelGAN and WaveRNN in our preliminary experiments.

V. EXPERIMENTS AND ANALYSIS

A. Dataset

MPop600 is a singing voice database compiled in our prior work [22]. It contains 600 Mandarin pop songs sung by 2 male and 2 female vocalists, along with their corresponding musical scores, which were semi-automatically transcribed. Each audio contains a single vocal without any background music, and only the first verse and the chorus were recorded. Within the scope of this paper, we built a SVS system that focuses on one vocalist; 150 songs sung by the same female singer (female 1 in Fig. 3) were utilized for the experiment. This subset contains about 3 hours of audio recorded at 96 kHz sampling rate with a resolution of 24 bits per sample, but were down-sampled to 22.05 kHz for the experiment. We chose 3 songs for validation and 2 songs for testing.

B. Experimental Conditions

In training the Tacotron2-based network for predicting mel-spectrograms and the PWG for audio generation, the hyper-parameters and training setups were set to be the same configurations as in [8] and [14], respectively. In this research, five models were constructed to evaluate the effect of duration models in SVS:

- **model 1**: without informing Tacotron2 of the phoneme duration,
- **model 2**: the LSTM-based duration model [5],
- **model 3**: the proposed rule-based duration model,
- **model 4**: model 3 with tonality consideration, and
- **model 5**: model 4 with data augmentation.

To verify the necessity of a duration model for SVS, model 1 is trained by feeding phoneme-level input sequences into a regular Tacotron2 without informing the phoneme duration [8]. Complying with the architecture in Fig. 5, model 2∼5 were established to compare the effect of different duration models in SVS. We trained model 2 with an LSTM-based duration model following the paradigm of prediction and post-processing set out by Kim *et al.* [5] as the baseline for comparing the performance of the proposed rule-based duration model, denoted as model 3.

In addition, since Mandarin is a tonal language, model 4 considered Mandarin tonality by adding a rule 0 before rule 1 in the algorithm, which has the same condition and implementation as in rule 1 but has to meet the requirement of the same tonality. This imposed a stronger restriction of estimating the phoneme duration of a stretched Mandarin syllable. Furthermore, model 5 applied data augmentation to stretch the audios in the training set by 1.5 times using iZotope RX7[4]. Tempos of the corresponding musical scores were also slowed down by 1.5 times, but phoneme durations were re-detected by forced alignment. As a result, this gave the training set more data with a variety of note lengths that can be matched with testing data in rule 1, which is the most promising rule in the algorithm.

C. Evaluation Items

To evaluate the accuracy of the fundamental frequency (F0) of the singing voice, F0 was extracted by a pre-trained CREPE pitch detector[5] [29] from the synthesized audio. After that, root-mean-square error of F0 (F0 $er.$) in the unit of semitone (use base-2 logarithm and multiplied by 12) and F0 correlation (F0 $cor.$) between the reference audio and synthetic voice were calculated. For fair comparisons against a performance upper bound, the *reference audio* was also generated via re-synthesis from the ground-truth mel-spectrograms directly using PWG.

[4]https://www.izotope.com/en/products/rx/features.html
[5]https://github.com/marl/crepe

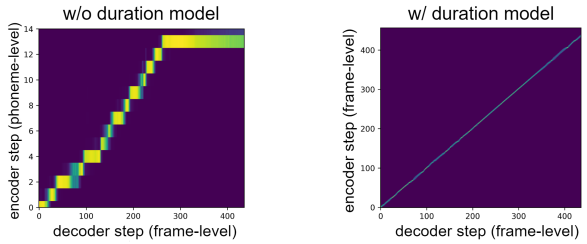

Fig. 7. Comparison on attention alignments w/ or w/o a duration model.

Additional objective metrics, including root-mean-square error and correlation of phoneme duration (Dur *er*. and Dur *cor*.), and mel-cepstral distortion (MCD), which is commonly used for synthesized speech quality assessment, are all presented in Table I.

Moreover, a listening test was carried out, and the mean opinion score (MOS) among the 15 subjects who participated in the evaluation are shown in Fig. 6. The participants were instructed to make their judgment based on the pitch and pronunciation accuracy of the generated voices and give an overall rating with a scale from 1 (poor) to 5 (good).

D. Evaluation on Overall Performance

The evaluation results in both Table I and Fig. 6 show that model 1 gave the worst performance due to the coarse-grained alignment of directly mapping from phoneme-level to frame-level (see the first panel in Fig. 7). In contrast, model 2∼5 exhibited the same trend illustrated in the second panel of Fig. 7 which demonstrates the attention alignment between frame-level encoder step and that of the decoder. The pre-expansion of features not only increases the model capacity for the richer information, but also enhances the precision and accuracy of alignment. It is significant to realize that the jeopardy of alignment mismatch is not just reflected in the tempo but the comprehensive quality of the synthesis since it impacts the decoded mel-sequence containing the information of F0, timbre, and pronunciation.

On top of that, the rule-based model outperformed the LSTM-based model in all aspects as shown in both objective and subjective evaluations. Fig. 8 compares the F0 contour generated by the reference audio, model 2, and model 3 for one particular singing example. It demonstrates that the rule-based duration model produced a pitch contour that was more consistent to that of the reference audio. These results confirm the effectiveness of our phonology-based duration prediction algorithm.

In addition, as Mandarin is a tonal language, we added the tonality consideration before rule 1 and observed improvements across the evaluation metrics of model 4. It verifies that the phoneme duration may depend slightly on the tonality. Furthermore, when we augmented the data by time stretching, the overall performance of model 5 improved both subjectively and objectively. This may be due to the availability of more exemplars that met the requirement of rule 1.

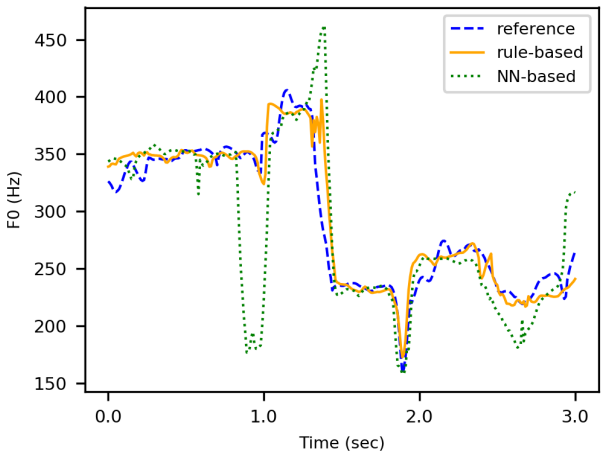

Fig. 8. Examples of F0 contour for SVS with LSTM duration model (NN-based) and the proposed duration model (rule-based), as compared against the reference audio.

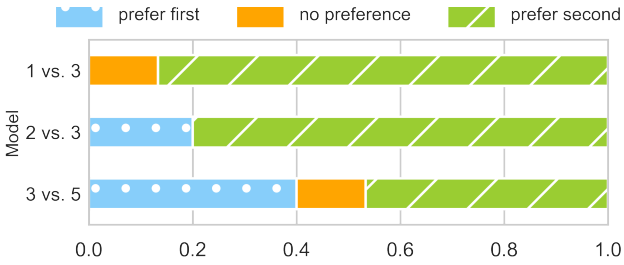

Fig. 9. Preference distribution for rhythm on each paired group.

E. Evaluation on Duration Model by Listeners' Preference

To examine the rhythmic performance, another subjective evaluation was conducted in the form of a blind A/B preference test, and the results are summarized in Fig. 9. The 15 subjects were requested to listen to the reference audio before comparing the same song segment generated by two different models, and select the one with a relatively more natural rhythm. The comparison consists of model 1 vs. 3 (duration informed or not), 2 vs. 3 (NN-based vs. rule-based), and 3 vs. 5 (more consideration on 3). It turned out that 80% supported model 3 rather than model 2. This preference result might be mainly because that the low predicted duration error in model 3 ensured the naturalness of perceived singing rhythm. The preference toward model 5 vs. model 3, however, did not come out as definite as suggested by the MOS.

VI. Conclusions

In this research, we established a rule-based duration model inspired by linguistic observations on Mandarin phoneme durations. This proposed model aimed to achieve a natural rhythm in SVS tasks even with small datasets. Both the literature review and our statistical analyses on the dataset

supported the proposed model's fundamental assumption that *initial ratios* in different stretched syllables should display little variation. With this principle, our duration prediction algorithm multiplies the note length with the predicted *initial ratio* to obtain the estimated phoneme duration. The duration information is incorporated through a length regulator into our end-to-end SVS system composed of a Tacotron2-based acoustic model and a PWG vocoder. In our experiment, phoneme durations obtained through forced alignment were used in training and predicted durations were used in inference. The experiment compared SVS models with different duration model setups and showed that the proposed rule-based model outperformed its NN-based counterpart comprehensively in terms of naturalness of rhythm, pitch, and pronunciation. Finally, the inclusion of the tonality consideration and data augmentation was shown to have enhanced the quality of the synthesized singing voice by providing fine-grained alignments between the encoded and decoded sequences.

ACKNOWLEDGMENT

This research is supported by the Ministry of Science and Technology (MOST) of Taiwan under Grant No. 108-2634-F-007-003 and No. 109-2221-E-007-094-MY3 awarded to SHW and YWL.

REFERENCES

[1] H. Kenmochi and H. Ohshita, "Vocaloid-commercial singing synthesizer based on sample concatenation," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2007.

[2] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-based singing voice synthesis system," in *International Conference on Spoken Language Processing*, 2006.

[3] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on deep neural networks," in *INTERSPEECH*, 2016, pp. 2478–2482.

[4] K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on convolutional neural networks," *arXiv* preprint arXiv:1904.06868, 2019.

[5] J. Kim, H. Choi, J. Park, M. Hahn, S. J. Kim, and J. J. Kim, "Korean singing voice synthesis based on an LSTM recurrent neural network," in *INTERSPEECH*, 2018, pp. 1551–1555.

[6] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6955–6959.

[7] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *INTERSPEECH*, 2017, pp. 4006–4010.

[8] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*, 2018, pp. 4779–4783.

[9] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *International Conference on Learning Representations (ICLR)*, 2018.

[10] Y. Gu, X. Yin, Y. Rao, Y. Wan, B. Tang, Y. Zhang, J. Chen, Y. Wang, and Z. Ma, "ByteSing: A chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and WaveRNN vocoders," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021.

[11] J. Lee, H. Choi, C. Jeon, J. Koo, and K. Lee, "Adversarially trained end-to-end Korean singing voice synthesis system," in *INTERSPEECH*, 2019, pp. 2588–2592.

[12] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv* preprint arXiv:1609.03499, 2016.

[13] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning (ICML)*, 2018, pp. 2415–2424.

[14] R. Yamamoto, E. Song and J. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*, 2020, pp. 6199-6203.

[15] J. Chen, X. Tan, J. Luan, T. Qin, and T. Liu, "HiFiSinger: Towards high-fidelity neural singing voice synthesis," *arXiv* preprint arXiv:2009.01776, 2020.

[16] J. Liu, C. Li, Y. Ren, F. Chen, P. Liu, and Z. Zhao, "Diffsinger: Diffusion acoustic model for singing voice synthesis," *arXiv* preprint arXiv:2105.02446, 2021.

[17] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech: Fast, robust and controllable text to speech," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 3165–3174.

[18] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, D. Su, and D. Yu, "Durian: Duration informed attention network for multimodal synthesis," *arXiv* preprint arXiv:1909.01700, 2019.

[19] M. Blaauw and J. Bonada, "Sequence-to-sequence singing synthesis using the feed-forward transformer," in *ICASSP*, 2020, pp. 7229–7233.

[20] Y. Wu, S. Li, C. Yu, H. Lu, C. Weng, L. Zhang, and D. Yu, "Peking Opera Synthesis via Duration Informed Attention Network," *arXiv* preprint arXiv:2008.03029, 2020.

[21] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, "XiaoiceSing: A high-quality and integrated singing voice synthesis system," *arXiv* preprint arXiv:2006.06261, 2020.

[22] C. Chu, F. Yang, Y. Lee, Y. Liu and S. Wu, "MPop600: A Mandarin popular song database with aligned audio, lyrics, and musical scores for singing voice synthesis," in *Asia-Pacific Signal and Information Processing Association (APSIPA)*, 2020, pp. 1647-1652.

[23] C. Lü, *Chinese literacy learning in an immersion program*. London: Palgrave Macmillan, 2019.

[24] Y. Lee, T. Liao, and Y. Liu, "A simple strategy for natural Mandarin spoken word stretching via the vocoder," in *International Congress on Acoustics (ICA)*, 2019.

[25] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," in *Speech Communication*, 1995, pp. 175–205.

[26] J. Driedger, "Time-scale modication algorithms for music audio signals," Master's thesis, Saarland Univ., 2011.

[27] Z. Duan, H. Fang, B. Li, K. C. Sim and Y. Wang, "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *APSIPA*, 2013, pp. 1-9.

[28] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *NeurIPS*, 2015, pp. 577–585.

[29] J. W. Kim, J. Salamon, P. Li and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *ICASSP*, 2018, pp. 161-165.