An Effective Initial/Final Duration Prediction Method for Corpus-based Singing Voice Synthesis of Mandarin Chinese

Cheng-Yuan Lin, Pei-Chi Jao and J.-S. Roger Jang

Department of Computer Science, National Tsing Hua University, Taiwan

{gavins, peggy, jang}@wayne.cs.nthu.edu.tw

Abstract

In this paper, we propose an effective method for predicting initial/final duration for corpus-based singing voice synthesis of Mandarin Chinese. The goal of the method is to improve the naturalness and clarity of the synthesized singing voices. To achieve this goal, we construct an individual initial/final (I/F) duration prediction model for each category of consonants. Support vector machine is used for duration prediction in each model. In order to achieve better accuracy, we use both linguistic/phonetic attributes and music-score information as the input features for the I/F duration prediction model. Experimental results demonstrate that the proposed method is effective in predicting the I/F duration for singing voice synthesis.

Index Terms: speech synthesis, singing voice synthesis, initial/final duration prediction, regression model.

1. Introduction

Singing differs significantly from speech in terms of its production and perception by human beings. In addition to intelligibility which is the major focus in speech, sound quality and musicality are also essential in singing. Due to wide variations in pitch and loudness, the singing voice has been difficult to be modeled and reproduced satisfactorily. Most previous approaches for singing voice synthesis have employed models that tried to characterize human speech production mechanism. For instance, the SPASM system developed by Cook [1] employed an articulator-based tube representation of the vocal tract and a time-domain glottal pulse input to approximate the articulatory system. Sinusoidal models are more general representations that are capable of high-quality modeling, modification, and synthesis of both speech and music signals [2]. Recently, the corpus-based approaches have been applied for singing voice synthesis (SVS). For example, a unit selection based on sinusoidal modeling, with a singing voice inventory of 500 nonsense words sung in several pitch levels was proposed by [3]. We also proposed a corpus-based singing voice synthesis system for Mandarin Chinese in the prior study [4].

Based on our observations, a perceived delay (or time-lag) may occur in the synthesized outputs, especially for some specific syllables. The perceived delay exists even if the onset of a syllable is aligned perfectly with the beginning of a note. Besides, some syllables may sound like others if the initial duration prediction is not satisfactory. For instance, the syllable 'ㄕㄚ' (sha) (see the left plot of Figure 1) would sound like the syllable 'ㄓㄚ' (zha) if we shorten the initial duration while keeping the same length of the syllable. (see the right plot of Figure 1). Due to the fact that the duration of each syllable is determined by its corresponding music score, there is no need to predict the duration of a syllable in SVS. Therefore, the goal of this study is to develop a reliable I/F duration prediction which estimates the I/F duration of a syllable when the duration of the syllable is specified. In this paper, we employ several influential features with an effective regression approach, support vector machine (SVM), to construct the prediction model. In order to validate the feasibility of the proposed method, several experiments are conducted accordingly. The details will be explained in the following sections.

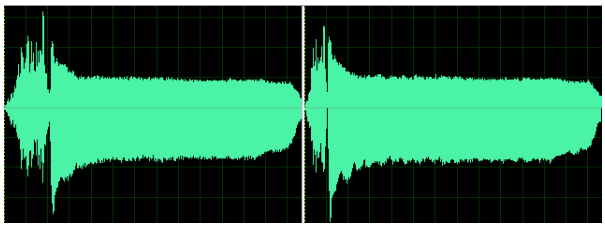

Figure 1: *left segment: the syllable 'ㄕㄚ' (sha) sung by a singer; right segment: the length of the initial part of the syllable 'ㄕㄚ' (sha) is shortened to 50% of its original length and it sounds like the syllable 'ㄓㄚ'(zha).*

This paper is organized as follows: Section 2 briefly describes our prior study concerning corpus-based SVS. Section 3 elaborates the framework of the initial/final duration prediction model. Section 4 presents the experimental results. Finally, Section 5 comes with the conclusions and future work.

2. Prior work review: a corpus-based singing voice synthesis system of Mandarin Chinese

In our prior work, three kinds of singing voice corpora are collected. They are single-syllable-based corpus (SSC), coarticulation-based corpus (CC), and song-based corpus (SC), respectively. Detailed descriptions of these corpora are given as follows.

- SSC consists of six types of recordings which are the combinations of three different pitch levels and two different durations for each Mandarin syllable.
- CC covers a total of 751 syllabic pairs which were collected based on the commonly occurred coarticulation in singing of Mandarin.
- SC is composed of 1384 non-uniform length sentences which are from 30 popular Mandarin songs.

A female professional dubber was requested to record the three corpora, in which a total of 12 742 syllables are collected. Subsequently, we carried out a series of preprocessing tasks, including initial/final segmentation by using a previously proposed automatic phonetic segmentation algorithm [5], pitch estimation by using a robust pitch tracking method [6], pitch marking based on a two-phase algorithm [7], etc. Although these tasks are processed automatically, in order

to ensure the reliability of the data for singing voice synthesis, human inspection was called for to avoid any possible errors.

In the synthesis processing, we used two distance functions based on dynamic programming to select the optimum units from the three corpora. The two distance functions are primarily designed to measure the differences of pitch and duration between target units and source units. The detailed descriptions can be found in our prior study [4].

3. Initial/final duration prediction model

Since the overall duration of a syllable is given according to the music score, all we need to do is scaling the duration of a syllable to its target length. Three ways to specify the I/F duration of a syllable are listed next.

- Keep the original length of the initial part of a syllable and only change the final part for the newly synthesized syllable.
- Use the ratio of the I/F durations of the original syllable for the newly synthesized syllable.
- Construct a prediction model to estimate the I/F durations of the newly synthesize syllable.

Based on our observations, neither the first nor the second method gave satisfactory performance in practice. Hence we adopted the third method for our system. In fact, such a prediction model is very common in TTS systems. It is thus intuitive to adopt the same framework to construct the I/F duration prediction model. In other words, a set of influential input features with the corresponding output features needs to be collected in advance as the training data set, then an effective regression approach is employed to construct the model. Details will be elaborated in the following sections.

3.1. The input features and output features for the I/F duration prediction model

In Mandarin singing voice synthesis, if the initial of a syllable is aperiodic (or unvoiced, ex. 'ㄕ' (sh)), we need to perform pitch-scale modification on its final at first, and then to perform time-scale modification on both its initial and final, respectively. Finally, we concatenate the two parts into a synthesized syllable. On the other hand, if both the initial and the final of a syllable are periodic, we perform pitch-scale and time-scale modifications directly on the syllable. In other words, there is no need to know the I/F durations during the synthesis when the initial and the final of a syllable are both periodic. For example, for syllable 'ㄇㄚ' (ma), we can alter its pitch and duration as a whole regardless of the boundary between the initial and final. In view of this, we classified all Mandarin consonants into two sets, periodic and aperiodic groups, according to the periodicity of their waveforms. The contents of the two groups are listed in Table 1.

Table 1. *Periodic/aperiodic groups of consonants*

Group	Consonants
periodic	m, n, l, r, 'null'
aperiodic	h, x, sh, j, zh, z, b, d, g, p, t, k, q, ch, c, f, s

In theory, we can construct only one prediction model for the aperiodic group. However, the variations among the acoustic characteristics of the consonants in the aperiodic group are usually too significant to ignore. Hence, a better way is to construct a prediction model for each aperiodic consonant. As a result, we have 17 kinds of I/F duration prediction models finally. Since the classification of these models is based on the categories of consonants, we thus choose the initial durations as the desired output values for these models.

In addition to the output values, we also have to prepare corresponding input features for each prediction model. As in most of TTS systems, it is quite common to use several linguistic and phonetic attributes (see Table 2) with a regression approach to predict the durations of initial and final. It is noted that the features related with word length presented in Table 2 are measured in syllables. Therefore, it is intuitive to employ the same procedure to estimate the I/F duration of a syllable in SVS. However, the I/F duration prediction of TTS is still somewhat different from that of SVS. For instance, the duration of a syllable is known for SVS, whereas it is unknown for TTS. Moreover, the duration of a syllable for SVS is determined by the music score instead of the speaker.

In view of this, employing linguistic and phonetic attributes is likely to be insufficient or ineffective to predict the I/F duration of a syllable for SVS. Fortunately, in addition to linguistic and phonetic attributes, the prosody of a syllable in SVS is also known due to the available music score. Accordingly, we added another set of attributes (see Table 3) induced from the music score. It is noted that each attribute shown in Table 3 is further normalized to have zero sample mean and unity sample variance in our implementation. Eventually, we chose a set of linguistic and phonetic attributes as well as the music-score based attributes as the input features for each prediction model. For simplicity in later discussions, linguistic and phonetic attributes are referred to as TTS features and the music-score based attributes are referred to as SVS features in the following sections. Figure 2 illustrates the schematic diagram of the I/F duration prediction.

3.2. Support vector machine for the I/F duration prediction model

Once both the input features and the output values are defined, then we need to select an effective regression approach. Here we compared three common regression approaches; they are linear regression (LR), classification and regression tree (CART) [9], and support vector machine (SVM) [10], respectively. Brief introduction to our use of these approaches is given below.

- LR attempts to model the relationship between input and output variables by fitting a linear equation to the observed data using the least-squares error criterion.
- CART is a tree-building technique. In our experiment, we need to know when to stop the split. First, we create a decision tree with each impure node containing ten or more observations to be split. Second, we adopt ten-fold cross validation to compute the cost for the whole data and estimate the best level of pruning. Finally, we prune

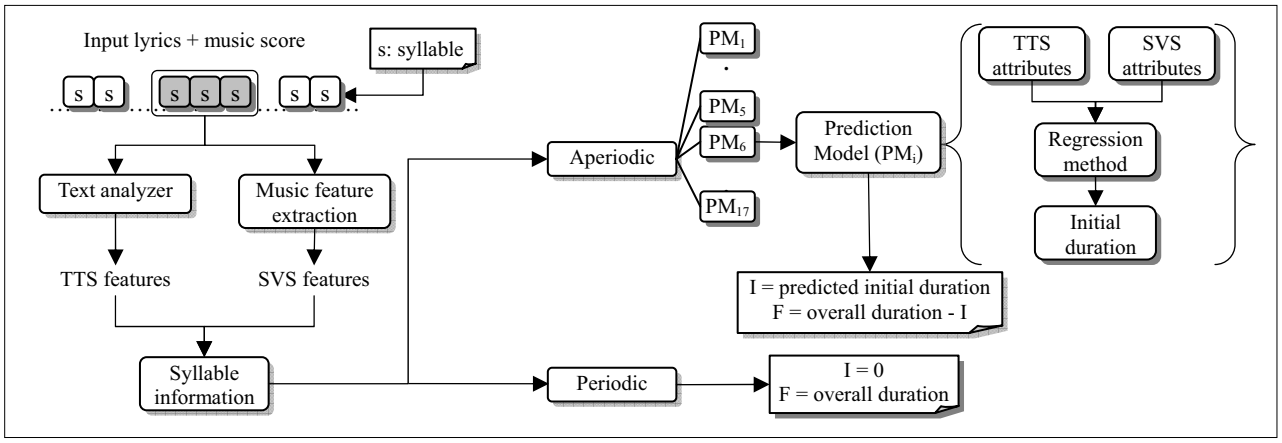

Figure 2: *Schematic diagram of the I/F duration prediction.*

Table 2. *Linguistic and phonetic attributes*

Attributes	Descriptions
$Vowel_c$	Category of vowel of the current syllable
$Vowel_p$	Category of vowel of the preceding syllable
$Tone_c$	Tone of the current syllable
$Tone_p$	Tone of the preceding syllable
$Tone_s$	Tone of the succeeding syllable
POS_c	POS (part-of-speech) of the lexical word that contains the current syllable
POSp	POS (part-of-speech) of the lexical word that contains the preceding syllable
POS_s	POS (part-of-speech) of the lexical word that contains the succeeding syllable
WL_c	Word length of the lexical word that contains the current syllable
WL_p	Word length of the lexical word that contains the preceding syllable
WL_s	Word length of the lexical word that contains the succeeding syllable
LWpos	Position of this syllable located in a lexical word
Spos	Position of this syllable located in a sentence

Table 3. *Music-score based attributes*

Attributes	Descriptions
$Duration_c$	The duration of the current syllable
$Duration_p$	The duration of the preceding syllable
$Duration_s$	The duration of the succeeding syllable
$Duration_{p-c}$	The duration difference between the preceding syllable and the current syllable
$Duration_{s-c}$	The duration difference between the succeeding syllable and the current syllable
$Pitch_c$	The average pitch of the current syllable
$Pitch_p$	The average pitch of the preceding syllable
$Pitch_s$	The average pitch of the succeeding syllable
$Pitch_{p-c}$	The pitch difference between the preceding syllable and the current syllable
$Pitch_{s-c}$	The pitch difference between the succeeding syllable and the current syllable

the results and get an optimal tree for each of the prediction model.

- The ε- support vector regression (ε-SVR) [11] and the ν-support vector regression (ν-SVR) [12] are the two common SVM methods used for regression. In this study, we chose ν-SVR and adopted the radial basis function (RBF) as its kernel function to map features into a higher dimensional space. In ν-SVR, two significant parameters, (C, γ) have to be determined. C denotes penalty parameter (a larger C corresponds to assigning a higher penalty to training errors) and γ denotes a weighting parameter used in the RBF kernel ($e^{-\gamma\|x_i - x_j\|^2}$). Here, we use a two-phase grid search with ten-fold cross validation [13] based on the initial configurations (C = $\{2^1, 2^2, \ldots, 2^{12}\}$ and $\gamma = \{2^{-15}, 2^{-9}, \ldots, 2^{-1}\}$) to evaluate the best ($C$, γ) pair for each prediction model.

Since the initial duration of a syllable is calculated by the regression model, its value could possibly be too large or too small. In order to find a reasonable value, we limit the predicted initial duration empirically with a lower bound and an upper bound via equation (1):

$$\frac{S_i}{3} \le Initial_i^{predicted} \le \min(3S_i, \frac{D_s}{2}) \quad (1)$$

where S_i denotes the standard deviation of the initial durations whose consonants belonged to the i^{th} category in the aperiodic group. D_s means the duration of a syllable which is determined by its corresponding music score. Once the initial duration of a syllable is estimated via the proposed method, we then employ waveform similarity overlap and add (WSOLA) method [14] to modify the I/F duration of a syllable.

4. Results and Discussions

In this section, we conducted several experiments to validate the feasibility of the proposed methods. First of all, we want to justify the use of SVS features in conjunction of TTS features. Two experiments using two feature sets of TTS features and TTS + SVS features were conducted respectively. Secondly, picking up a good regression approach is essential to the performance of the prediction models. Therefore, we adopted three regression approaches (LR, CART, and SVM) to compare their performance.

In addition to the experimental setups mentioned above, we selected the corpus SC (see Section 2) to carry out these experiments. We split the corpus into equal-size training and test sets. Root mean squared error (RMSE), and correlation (Corr) are used in performance evaluation.

Table 4. *Performance comparison of three regression methods using two feature sets in predicting initial duration. (Top: closed test, bottom: open test).*

Closed test (using training data)				
Regression Method		LR	CART	SVM
TTS features	RMSE	23.24	23.12	21.66
	Corr	0.888	0.889	0.904
TTS & SVS features	RMSE	21.31	20.83	**18.01**
	Corr	0.907	0.911	**0.935**
Mean		66.0115		
Standard deviation (STD)		48.8784		
Open test (using test data)				
Regression Method		LR	CART	SVM
TTS features	RMSE	23.94	24.34	23.06
	Corr	0.873	0.868	0.882
TTS & SVS features	RMSE	22.84	24.18	**21.58**
	Corr	0.885	0.871	**0.898**
Mean		65.1408		
Standard deviation (STD)		50.5475		

The unit of RMSE, Mean and STD is millisecond. It can be seen from Table 4 that using both TTS and SVS features based on SVM to construct the I/F duration prediction models can achieve the optimum performance among all the other combinations. Compared with the results by using TTS features alone in the I/F duration prediction, using both TTS and SVS features to predict the I/F durations indeed reduced the RMSE in both the closed and open test on three regression approaches mentioned above. The experimental results also implied that SVS features are definitely helpful to improve the prediction accuracies of the I/F durations. As a result, we employed TTS and SVS features based on SVM to construct the I/F duration prediction models in our SVS system finally.

In order to validate the practicality of the selected features/model further, we conducted a listening test. As noted in Section 3, we can use at least three methods to modify the I/F duration of a syllable. However, the first method poses a difficulty in implementation when the duration of the synthesized syllable is less than that of the initial part of the original syllable. Such situation hardly occurs in TTS systems, whereas it is likely to happen in SVS systems due to the fact that the variation of duration is large. In view of this, we only compared the other two methods in this study. For method 3, we adopted the proposed framework with TTS and SVS features and SVM regression.

Table 5. *Listening test using two different duration modification methods*

Sentence No / Methods	1	2	3	4	5	6	7	8
2nd (ballots)	3	1	2	2	0	2	3	1
3rd (ballots)	12	14	13	13	15	13	12	14

In this study, eight test sentences were randomly selected from the test data which was used in the previous open test experiment. Each sentence has two synthesized outputs generated by methods 2 and 3 (see Section 3), respectively. 15 subjects were invited to conduct the experiment. These subjects are native speakers of Mandarin Chinese, with an average age of 24. Each subject was asked to select for the one with better synthetic quality between two synthesized outputs for each test sentence according to his/her preference. It is noted that we put the two kinds of synthesized outputs in a random order. Since the listening test was conducted through a web page, they can listen to the stimuli as many times as they wanted before making a decision. Table 5 demonstrates the experimental results, where most of the subjects preferred the synthesized outputs generated by the proposed method. It also indicates that using the proposed I/F duration prediction model can actually improve the output sound quality in SVS.

5. Conclusions and Future Work

In this paper, we have presented the I/F duration prediction model for corpus-based singing voice synthesis of Mandarin Chinese. Under the framework of the proposed model, both TTS and SVS features are used as the input features for each I/F duration prediction model and SVM is chosen as the regression kernel of the model. In order to verify the feasibility of the proposed model, we conducted several experiments based on different regression approaches and feature sets. Furthermore, we also conducted a listening test to validate the effectiveness of the proposed method. The results demonstrate satisfactory performance of the proposed model. It is still possible to improve the efficiency and effectiveness of the I/F duration prediction further. For example, we can use the stepwise regression technique [15] to find out the most influential attributes among TTS and SVS features to reduce the unnecessary computation. Besides, we will try to simulate the actual pitch contours sung by a singer to perform the pitch prediction based on the similar ideas proposed in this study.

6. References

[1] P. R. Cook, "SPASM, a real-time vocal tract physical model controller and singer, the companion software synthesis system," *Computer Music Journal.*, Vol. 17, pp. 30-43, 1993.

[2] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech and Audio Proc*., Vol. 5, pp. 389-406, 1997.

[3] M. W. Macon, L. Jensen-Link, J. Oliverio, M. Clements, and E. B. George, "A singing voice synthesis system based on sinusoidal modeling," in *Proc. ICASSP,* 1997, pp. 435-438.

[4] C. Y. Lin, T. Y. Lin, and J. S. Jang, "A corpus-based singing voice synthesis system for Mandarin Chinese," in *Proc. ACM Multimedia*, 2005, pp. 359–362.

[5] F. C. Chou, C. Y. Tseng, and L. S. Lee, "Automatic segmental and prosodic labeling of Mandarin speech," in *Proc. ICSLP*, 1998, pp.1263-1266.

[6] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, 1995, pp. 495–518.

[7] C. Y. Lin and J. S. Jang, "A two-phase pitch marking method for TD-PSOLA synthesis," in *Proc. ICSLP*, 2004, pp. 1189-1192.

[8] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, 1998, pp.121-167.

[9] L. Breiman, *Classification and Regression Trees*, Chapman & Hall, Boca Raton, 1993.

[10] C. C. Chang and C. J. Lin, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[11] V. Vapnik, *Statistical learning theory*, Wiley, New York, 1998.

[12] B. Schölkopf, A. Smola, R. Williamson, and P. L. Bartlett, "New support vector algorithms," in *Proc. Neural Computation*, 2000, pp. 1207–1245.

[13] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Technical Report, Department of Computer Science & Information Engineering, National Taiwan University, Taiwan.

[14] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity for high quality time-scale modification of speech," in *Proc. ICASSP,* 1993, pp. 554-557.

[15] P. T. Pope and J. T. Webster, "The use of an f-statistic in stepwise regression procedures," *Technometrics*, 1972.