10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Segment Duration and Proportion in Mandarin Singing

Cong Zhang[1], Xinrong Wang[2]

[1]School of European Culture and Languages, University of Kent
[2] Department of Psychology and Language Sciences, UCL
c.zhang-74@kent.ac.uk, wxrsophie@outlook.com

Abstract

Speech-based singing synthesis has various merits while it also has unsolved issues. One of the most noticeable issues is the segment duration and proportion in synthesised singing, which is caused by the difference in the short syllables in speech and the lengthened syllables in singing. This study, therefore, investigates how syllables are lengthened in Mandarin singing data. A total of 20 songs from the MIREX singing corpus were segmented and analysed. The results showed that (1) the segment proportions in Mandarin syllables are different in speech and in singing; (2) the lengthening is influenced more by the slots in the syllable structure than by the types of segments; (3) in Mandarin, the rime nucleus lengthens the most and the rime codas follow. The durations of the vowel-like onsets also increase but their proportions in a syllable decrease.

Index Terms: singing synthesis, speech-sing interaction, language technology

1. Introduction

The recent trend in synthesising singing voice has brought about challenges in studying the phonetics of singing. Speech-based parametric singing synthesis is a popular adaptation of the speech synthesis method for singing. Using speech data to train singing models has various advantages over using singing data to train singing models: firstly, from the perspective of technological development, this method allows singing synthesis, aka text-to-sing, to build on the existing speech synthesis methods which are mature and controllable; secondly, from the perspective of data acquisition, speech-based singing synthesis is more economical since precise and clean singing data are hard to acquire and the annotation of singing poses yet another challenge; thirdly, this method also empowers more voices to sing, since only by providing speech data, singing voices can be synthesised. However, despite the good overall synthesis performance, a few types of syllables in Mandarin singing synthesis are often perceived as unnatural. The unnaturalness stems from a range of articulatory differences between singing and speaking, such as the formants, phonation, and duration e.g. [1], [2]. One of the most outstanding issues in synthesised singing is the distortion of segment proportions in a syllable, which is due to the lack of lengthened syllables in the model input speech data – syllables in singing are often lengthened to fit the melodic rhythm of the songs, which can be as long as 10 seconds in extreme cases. Previous studies show that the vowel proportion in Mandarin speaking data ranges from 51.6% to 74.6% [3]. Duplicating the same proportion in singing synthesis means some consonants need to be lengthened to 2.5s - 5s in a 10-second syllable, which is physically unachievable. It is, therefore, crucial to examine the proportions of each part of the syllables in singing data.

Currently, very few studies have looked into this issue. [4] analysed 20 English songs and their counterparts in speaking, which concludes that different types of consonants in English have different singing-to-speaking segment duration ratio: stops have the lowest ratio of approximately 1.4, while liquids have a much higher ratio of 2.24. They also reported the average consonant proportion in syllables: expectedly, semivowels and stops take up the largest and smallest proportions of approximately 30% and 20% respectively. Furthermore, depending on the syllable structures and the position of the consonants, the variance is substantial. For instance, affricates vary in taking up from 5% to 30% of the syllables when preceded by vowels or at the syllable-ending position. However, there has not been any similar studies on Mandarin Chinese so it is difficult to model the temporal relationship between singing and speaking in Mandarin. The current study therefore examines the segment durations and their proportions in Mandarin singing data. The major goal of this preliminary study is to provide the text-to-sing modelling algorithms with a linguistic analysis of singing data so as to improve the naturalness of synthesised singing.

2. Current Research

The differences in syllable structures between English and Chinese make the results of English phoneme proportions in [4] not applicable in Mandarin speech-based singing synthesis. In Mandarin, the syllable structure is often referred to as CGVX (Consonant – Glide – Vowel – Nasal/Vowel), with the minimal structure being a single V [5, pp. 71, 79]. It can also be represented as ONC (Onset – Nucleus – Coda) [5, p. 82]. The rimes consist of an obligatory nucleus and a possible coda (either a nasal consonant or the second part of a diphthong). (1) illustrates the mapping between the two structures.

(1) O N C
 ⋀ | |
 CG V X

In this study, we examine four types of syllables that performed exceptionally bad in Mandarin speech-based singing synthesis. Our general questions are: How do different parts of a syllable vary temporally as the syllable length varies? In Mandarin singing, are segment durations and their proportions in the syllables encoded in the skeletal slots or in the segment classes?

The first category is syllables with **approximant rimes** (as V in CGVX and NC in ONC). This category features two syllabic approximants [ɪ] and [ɻ̩] ([5], [6]) as the rime, which are preceded by the consonant onsets *z* [ts] / *c* [tsʰ] / *s* [s], and *zh* [tʂ] / *ch* [tʂʰ] / *sh* [ʂ], respectively. The syllabic approximants are considered as apical vowels by some researchers (see review in [7]). How the syllabic approximants are lengthened

10.21437/SpeechProsody.2020-122

proportionally when the syllables lengthen is the first question we investigate in this study (RQ1).

The second category is syllables with **approximant onsets** (*y* [j] and *w* [w]). These glides act as the C or G in CGVX ([5, p. 79]) and O in the ONC. The approximant consonants are often considered as semivowels or vowels. How do approximant onsets lengthen in Mandarin singing? This is the second question we address in this research (RQ2).

The third category is syllable rimes with **nasal codas** (*n*[n] and *ng*[ŋ]), i.e. when the C in ONC is a N which corresponds to the X in CGVX. This type involves nasal codas succeeding the rime nuclei which are also argued to be nasalised vowels without real nasal codas [8, p. 89]. In Mandarin singing, it is unclear which part of the rime is lengthened and how much they lengthen respectively. RQ3 thus examines the proportions of nuclear V and the coda N in the syllable rimes.

The last category is syllable rimes with **vowel codas** (e.g. *ai* [aɪ̯], *ou* [oʊ̯]), i.e. when the C in ONC is the second half of a diphthong, which also corresponds to the X in CGVX. [9] suggested that the ratio of the first part to the second part in diphthongs, such as *ai* [aɪ̯] and *ei* [eɪ̯], is approximately 6:4 in Mandarin speech; while for GV rimes, such as *ia* [ja] and *ue* [ye], the ratio is 4:6. RQ4 therefore investigates the proportions of the nuclear V1 and the coda V2 in the syllable rimes in Mandarin singing data.

The target segments in the first two types share the same segment class (approximant) but are in different slots, while the targets in the latter two types are in the same slot but belong to different classes.

3. Methods

3.1. Materials

We used 20 songs (mean length: 203.7s) sung by 20 different amateur singers from the MIREX 2018 Mandarin pop song dataset [10]. All the singers were native speakers of Mandarin and sang in Mandarin for this corpus. Among the 20 songs, 15 were clean sound files while the other five had traces of background music. The corpus consists of wav files with a sampling rate of 44.1kHz, 16-bit, mono channel.

3.2. Procedures

Phonemes were first automatically segmented using Penn Phonetics Lab Forced Aligner for Chinese [11] and then manually checked, following these principles: (1) making the decision based on a general consideration of both the waveform and the spectrogram; (2) cutting at the mid-point of the transitional period between two segments if the transitional period of the two segments were long; (3) only altering the alignment when the automatic segmentation had clear mistakes and could be consistently corrected manually.

3.3. Data Analysis

For statistical analysis, simple linear regression models were built with dependent variables (DV) as predicted variables, and independent variables (IDV) as predictors. The model was: DV = (intercept) + (slope) * IDV. The DVs in this study are the syllable durations (for RQ1 and RQ2) or syllable rime durations (for RQ3 and RQ4). The IDVs in this study are specified in each section of the results according to their respective research questions. We also ran more complicated models such as polynomial models, spline models, and mixed effect model; however, simple linear regression models describe the data best by presenting the highest R^2 values.

4. Results and Discussions

4.1. RQ1: Approximant rimes

Table 1 shows the mean length of the Cs (consonants), the As (i.e. approximants), and their corresponding syllables. The maximum syllable length stretches to 2.11s while the longest consonant only reaches 0.28s.

Table 1: RQ1 – Durations of segments and syllables (sec)

	N		mean	sd	max	min
C	284	seg	0.12	0.04	0.28	0.03
		syl	0.42	0.28	2.11	0.08
A	284	seg	0.30	0.27	1.91	0.03
		syl	0.42	0.28	2.11	0.08

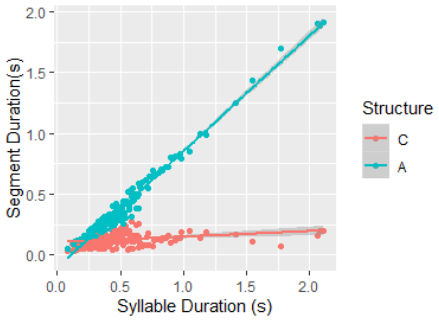

Figure 1: linear models for segment duration (RQ1)

The linear regression models were built for the durations of the onsets and the rimes separately to capture their different slopes and intercepts. These models are illustrated in Figure 1, in which the As have a much sharper slope than the Cs. The models showed a significant relationship between rime duration and syllable duration ($t = 109.64$, $p < 0.001$, $R^2 = 0.977$), onset duration and syllable duration ($t = 5.246$, $p < 0.001$, $R^2 = 0.089$).

The slope coefficient for rime duration was 0.954 for syllable duration, which indicates the rime segments decrease by 0.954s for each extra second of syllable duration. Summarising all slopes and intercept values, we have the following linear models:

- consonant onsets: 0.102 + 0.046 × Syllable duration
- approximant rimes: - 0.101 + 0.954 × Syllable duration

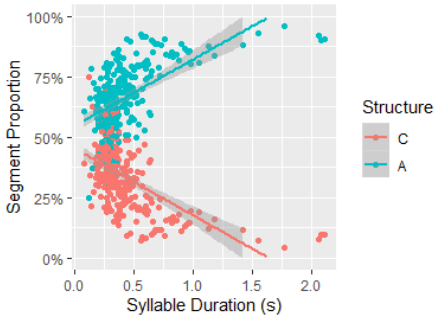

Figure 2: linear models for segment proportion (RQ1)

Since the regressions of durations are not completely the same as segment proportion, Figure 2 presents two other linear models for the consonant onset proportion and the approximant rime proportion in relation to syllables. The models showed a

significant relationship between rime proportion ($t = 12.09$, $p < 0.001$, $R^2 = 0.341$), onset proportion ($t = -12.1$, $p < 0.001$, $R^2 = 0.342$). Summarising all slopes and intercept values, we have the following linear models:

- consonant onsets: 45.3% - 0.276 × Syllable duration
- approximant rimes: 54.8% + 0.275 × Syllable duration

The results present a clear distinction between how the consonant onset and syllabic approximant rimes change differently. The rimes have a much sharper rise than the consonants. The consonants only have a small increase as the syllable is lengthened. The proportional change between the onsets and the rimes is also clear: the longer the syllable, the lower the percentage of the onset. Although the rimes are syllabic approximants, when the syllables are lengthened in singing, the rimes behave like full vowels. The results support that the skeletal slots encode the segment duration in Mandarin singing.

4.2. RQ2: Approximant onsets

The descriptive data in Table 2 show that the approximant onset is shorter than the vowels and nasals in the syllable structures of AVN and AV1V2.

Table 2: RQ 2 – Durations of segments and syllables (sec)

structure	N		mean	sd	max	min
N	239	seg	0.172	0.137	0.920	0.030
		syl	0.523	0.35	2.11	0.11
V	565	seg	0.293	0.311	4.480	0.030
		syl	0.457	0.357	4.53	0.0811
V1	229	seg	0.282	0.370	2.990	0.040
		syl	0.483	0.46	3.69	0.13
V2	229	seg	0.107	0.143	1.220	0.022
		syl	0.483	0.46	3.69	0.13
A	793	seg	0.092	0.053	0.400	0.022
		syl	0.465	0.389	4.53	0.081

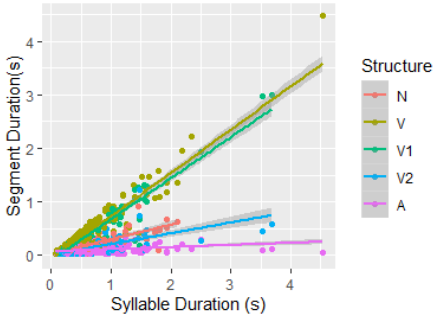

Figure 3: linear models for segment duration (RQ2)

Figure 3 shows the linear models of the approximant onsets duration against syllable duration as well as other components of the rimes against the syllable. Approximants act more like the consonant onset in the previous research question, while V2 (i.e. the second half of a diphthong) and N (nasal coda) have a slightly sharper slope. The V and V1, which are both in syllable nucleus position, have similar sharp slopes. The approximant onsets also significantly correlate with the syllable ($t = 8.495$, $p < 0.001$, $R^2 = 0.084$). The linear model for the approximant onsets is:

- approximant onsets: 0.073 + 0.039 × Syllable duration

Compared with the consonant onset model in RQ1 results, the slope of approximants (0.039) is even gentler than that of the affricates/fricatives in the previous part (0.046), which suggests the durations of the approximant onsets differ very little no matter how long the syllables are.

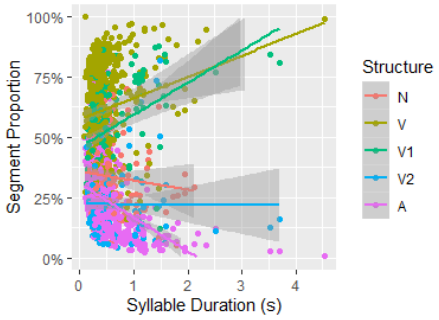

Figure 4: linear models for segment proportion (RQ2)

The segment proportion data in Figure 4 show that the approximant onsets have a falling slope as the syllables lengthened. The proportions of V2 and N are also falling but not as dramatically. The nucleus vowels take up more and more proportion as the syllables become longer. The approximant onsets duration model is as follows ($t = -13.72$, $p < 0.001$, $R^2 = 0.192$):

- approximant onsets: 30.1% - 0.141 × Syllable duration

In this part of the study, the approximants are more consonant-like than the consonant onsets in the previous part although they are approximants. The results here further support our argument that the slots frame the duration proportions.

4.3. RQ3 & RQ4: Nasal codas & vowel codas

Table 3 presents the results for both RQ3 and RQ4, since both of them have a structure of VX in the rime. The descriptive results in the table suggest that the VX are comparable since both V1 and V are the rime nuclei of the syllables, and V2 and N are the rime codas.

Table 3: RQ3 & RQ4 – Descriptive results of the duration of segment and rime (seconds)

structure	N		mean	sd	max	min
V1	1572	seg	0.287	0.304	3.300	0.020
		rim	0.416	0.379	3.88	0.036
V2	1572	seg	0.129	0.152	1.640	0.012
		rim	0.416	0.379	3.88	0.036
V	1812	seg	0.243	0.284	3.960	0.030
		rim	0.405	0.37	4.38	0.06
N	1806	seg	0.161	0.158	2.190	0.018
		rim	0.405	0.37	4.38	0.06

Figure 5 and Figure 6 display the linear models for the segment durations against syllable rime durations. The linear models again showed a significant relationship between V nucleus duration in VN rime ($t = 97.06$, $p < 0.001$, $R^2 = 0.839$), N coda duration in VN rime ($t = 40.83$, $p < 0.001$, $R^2 = 0.480$). The models are as follows:

- V nucleus: - 0.042 + 0.704 × Syllable duration
- N coda: 0.041 + 0.296 × Syllable duration

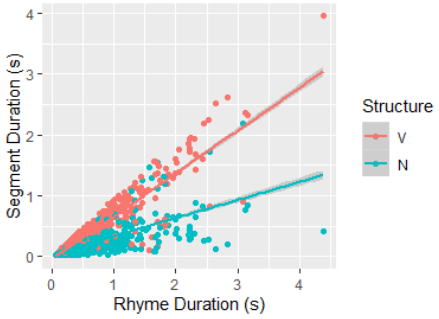

Figure 5: linear models for segment duration (RQ3)

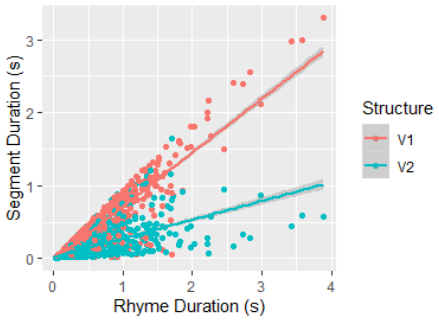

Figure 6: linear models for segment duration (RQ4)

The results from the diphthong data show very similar results, with V1 nucleus and V2 coda in VV rime matching with the V and N in VN rime. V1 duration is significantly correlated with the rime duration ($t = 92.410$, $p < 0.001$, $R^2 = 0.8447$), and so is V2 ($t = 32.887$, $p < 0.001$, $R^2 = 0.408$). The linear models are:

- V1 nucleus: -0.020 + 0.738 × Rime duration
- V2 coda: 0. 022 + 0. 257 × Rime duration

The slopes for V1 and V2 in the diphthongs follow the same pattern as the slopes for V and N in VN respectively. These results further supported the previous conclusions about the influence of syllable structure on the way segments lengthen.

The proportion data (Figure 7 and 8), however, exhibit very low R^2 value, although the models are statistically significant. The V's proportion in VN rime is significantly correlated with the rime duration ($t = 7.037$, $p < 0.001$, $R^2 = 0.027$), and so is N ($t = - 6.99$, $p < 0.001$, $R^2 = 0.026$). The linear models are:

- V nucleus: 54.3% + 0.078 × Rime duration
- N coda: 45.6% - 0.077 × Rime duration

In V1V2 rime, V1's proportion is correlated with the rime duration ($t = 2.283$, $p = 0.023$, $R^2 = 0.003$) and V2 is also marginally correlated ($t = -2.585$, $p = 0.00983$, $R^2 = 0.004$). The linear models are:

- V1 nucleus: 66.7% + 0.030 × Syllable duration
- V2 coda: 33.4% - 0.034 × Syllable duration

The major difference between VN and V1V2 is that, in V1V2, the change in the proportion is very subtle (both slopes are merely around 0.03), while V and N in VN have a sharper slope (0.078 and 0.077 respectively) than V1 and V2. It is not surprising since the diphthongs (V1V2) has a more stable structure than a VN rime.

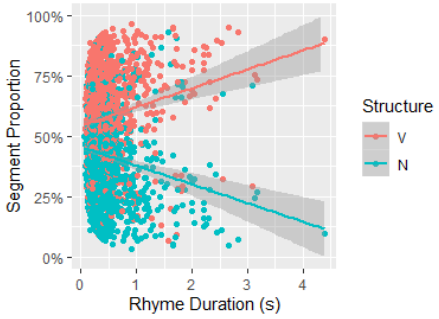

Figure 7: linear models for segment proportion (RQ3)

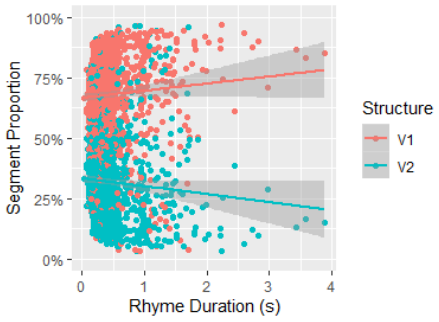

Figure 8: linear models for segment proportion (RQ4)

5. Conclusions and Limitations

The results of this study show that the skeletal slots significantly influence the duration in Mandarin singing, while the segment class seems to be a secondary predictor. In the ONC (Onset – Nucleus – Coda) structure of Mandarin syllables, we observe the following trends:

- the proportion of O falls substantially when a syllable or a rime lengthens;
- the proportion of N always rise as the syllable or the rime lengthens;
- the proportion of C falls slightly when a syllable or a rime lengthens.

The simple linear regression models in this study can explain the majority of the segment durations and how they change with the syllable duration; however, for the proportions of different segments, linear regression models are significant but can only explain a small portion of the data.

Since this is a preliminary study, there are many possibilities for future investigations on this topic: would the linear regressions change when speech-like syllable durations are excluded and only singing-specific syllables (i.e. long syllables) are modelled[1]? Are there more fine-grained factors that are influencing the model? Would the naturalness of the text-to-sing output improve providing the knowledge of syllable proportions? Can this study relate to speech rhythm studies and how? Future studies can benefit from a larger singing corpus with data of higher quality; matching speech corpus to cross-examine; more suitable modelling methods; and so forth.

[1] We would like to thank one of our anonymous reviewers for raising this question.

6. References

[1] J. Sundberg, "Articulatory differences between spoken and sung vowels in singers," *Stl-Qpsr*, vol. 10, no. 1, pp. 33–46, 1969.

[2] E. D. Bradley, "A comparison of the acoustic vowel spaces of speech and song," *Linguist. Res.*, vol. 35, no. 2, pp. 381–394, 2018, doi: 10.17250/khisli.35.2.201806.006.

[3] Y. Jue and D. Gibbon, "Criteria for database and tool design for speech timing analysis with special reference to Mandarin," in *Proceedings of the 2012 International Conference on Speech Database and Assessments, Oriental COCOSDA 2012*, 2012, pp. 41–46, doi: 10.1109/ICSDA.2012.6422453.

[4] 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 *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013, pp. 1–9, doi: 10.1109/APSIPA.2013.6694316.

[5] S. Duanmu, *The Phonology of Standard Chinese*, 2nd ed. Oxford University Press, 2000.

[6] S. I. Lee-Kim, "Revisiting Mandarin 'apical vowels': An articulatory and acoustic study," *J. Int. Phon. Assoc.*, vol. 44, no. 3, pp. 261–282, 2014, doi: 10.1017/S0025100314000267.

[7] Y. Xu and M. Wang, "Organizing syllables into groups-Evidence from F0 and duration patterns in Mandarin," *J. Phon.*, vol. 37, no. 4, pp. 502–520, Oct. 2009, doi: 10.1016/j.wocn.2009.08.003.

[8] S. Feng, *Prosodic syntax in Chinese: Theory and facts*. Routledge, 2019.

[9] J. Cao and S. Yang, "An experimental investigation on diphthongs in Standard Chinese," *Zhongguo Yuwen*, no. 6, pp. 426–433, 1984.

[10] R. Gong and G. Dzhambazov, "MIREX 2018 Mandarin pop song dataset," 2018. [Online]. Available: https://www.music-ir.org/mirex/wiki/2018:Automatic_Lyrics-to-Audio_Alignment#MIREX_2018_Mandarin_pop_song_dataset. [Accessed: 03-May-2019].

[11] J. Yuan, N. Ryant, and M. Liberman, "Automatic phonetic segmentation in Mandarin Chinese: boundary models, glottal features and tone," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2539–2543.