Synchronising speech segments with musical beats in Mandarin and English singing

Cong Zhang[1], Jian Zhu[2]

[1]Radboud University, Netherlands
[2]University of Michigan, USA

cong.zhang@ru.nl, lingjzhu@umich.edu

Abstract

Generating synthesised singing voice with models trained on speech data has many advantages due to the models' flexibility and controllability. However, since the information about the temporal relationship between segments and beats are lacking in speech training data, the synthesised singing may sound off-beat at times. Therefore, the availability of the information on the temporal relationship between speech segments and music beats is crucial. The current study investigated the segment-beat synchronisation in singing data, with hypotheses formed based on the linguistics theories of P-centre and sonority hierarchy. A Mandarin corpus and an English corpus of professional singing data were manually annotated and analysed. The results showed that the presence of musical beats was more dependent on segment duration than sonority. However, the sonority hierarchy and the P-centre theory were highly related to the location of beats. Mandarin and English demonstrated cross-linguistic variations despite exhibiting common patterns.

Index Terms: singing analysis, singing voice synthesis, rhythm, P centre, sonority

1. Introduction

The world-wide success of virtual singers, such as Hatsune Miku, has generated widespread interest in further development and research into synthesised singing. Over the course of the past few decades, the advancement of text-to-speech algorithms has immensely improved the performance of synthesised singing. The state-of-the-art models make use of deep learning methods (e.g. [1, 2]), which generate much more natural singing voice than the traditional concatenative or parametric models. However, deep learning models usually require more resources to build and to run. These shortcomings suggest that the lightweight and flexible traditional parametric methods are still useful in specific settings. Among all alternatives, optimising existing text-to-speech models trained with speech data for the use of singing voice synthesis seems to be still relevant.

Admittedly, compared with deep learning models, speech-based models generate singing voices that are less natural; they nevertheless allow a wider community to access the resources and create their own singing voice. The majority of the issues exist due to the lack of alignment between musical information and speech information. Among all issues induced by the missing information, the temporal alignment between speech segments and musical beats strikes as the most serious issue: the synthesised singing sounds off-beat at random and thus sounds unnatural to human ears. This is because musical beat information is not present in the speech training data, and therefore cannot be acquired through machine learning. This information thus needs to be provided to the models in other ways. Previous studies have adopted various methods in resolving this issue. [3, 4] used a heuristic way of assigning duration values or fittings to different segments (e.g. setting one phoneme to 20 ms, and another to 100 ms). [5] proposed an HMM-based "time-lag" model that was trained on 60 Japanese songs to account for the temporal difference between the musical notes and phonemes. With the addition of the time-lag model, the mean opinion score (MOS) was substantially improved. In more recent studies, the importance of duration was also highlighted and modelled using various deep learning models (e.g. [4, 6, 7]). However, these duration models still required a substantial amount of data for model training. This study aims to deliver an interdisciplinary analysis using linguistic and music knowledge, to inform future singing voice synthesis model building. The results can, for instance, replace the heuristic values used in such studies as [3, 4]. By providing a way of enhancing the temporal features, this study can also be beneficial to the more complicated modelling methods. It can also supplement the current modelling algorithms with domain knowledge, without attempting to fit any predictive synthesis models. An additional goal of this study is to gain insights into speech rhythm through analysing singing data that contain speech segments and fixed rhythm.

1.1. Linguistics Background

Speech Rhythm. While speech is not guided by musical scores, it encodes a natural rhythm. Many past studies have examined how speech rhythm was produced and perceived. One of the most relevant theories is the "P-centre" theory, i.e. perceptual centre. It was proposed to account for the alignment of timing units and speech segments [8]. Both acoustic and perceptual studies (e.g. [9, 10, 11, 12]) have found the factors influencing the locations of P-centres to be the phonological structure of syllables. In a syllable consisting of a syllable-initial consonant and a subsequent vowel, the P-centre is usually close to the vowel onset [13, 14, 15]. [15] has also found the energy profile of syllable-initial consonants to be important to the P-centre location. Many other traditional speech rhythm metrics such as $\%V$, ΔV and so forth are also tied to the timing of vowels. Recent studies employed more up-to-date metrics such as `maxE` and `maxD`, and also concluded the importance of vowel onsets and energy in the synchronisation of speech rhythm [16].

Sonority Hierarchy. Another relevant topic is the sonority hierarchy of speech segments [17]. Many studies have looked into the sonority of different segment types, and the most widely agreed version of the hierarchy is: `vowels > glides > liquids > nasals > obstruents` [18]. While the more sonorant types (vowels, glides, liquids, nasals) attracted more agreement, the order of the three obstruent types (stops, fricatives, and affricates) has been much more controversial.

[19] studied the acoustic correlates and found that while intensity contributed to the majority part of the sonority hierarchy, other parameters such as formant, duration, and peak air flow all served as relevant factors. Sonority hierarchy is largely universal; however, minor cross-linguistic variations also exist.

1.2. The Current Study

We analysed a Mandarin singing corpus and an English singing corpus. The two languages were chosen because, according to tradition rhythmic class theories, Mandarin and English were considered as syllable-timed and stressed-timed respectively [20, 21]. While we have reservations about the theoretical validity of this classification, it offers a convenient division to account for the typological differences between the two languages in terms of syllable structures, vowel reduction rules, etc. The following research questions and hypotheses were addressed:

1. What types of speech units attract most of the beats? Since we expect the alignment to favour vowels, it is likely that other sonorant phoneme types also anchor musical beats. We hypothesise that the more sonorant a segment is, the more likely it is to have beats aligned with it.

2. Where do musical beats align with speech segments? The theories related to speech rhythm such as P-centre theory suggest that the vowel onsets and the transition between the preceding consonants and the vowels are more likely to be the synchronising points for speech rhythm. We therefore hypothesise that the rimes in Mandarin and the vowels in English are aligned with the most beats, and the beat locations are likely to be at the beginning of the vowels or the end of preceding consonants. Moreover, the more sonorant the segments are, the more vowel-like they are, and thus the earlier the alignments are.

3. Are there cross-linguistic differences? We expect Mandarin and English to exhibit different synchronisation patterns. Given that sonority hierarchies can exhibit language-specific variations, and English and Mandarin belong to different linguistic typologies in many respects, the two languages are likely to exhibit speech-beat synchronisation differences.

2. Mandarin Singing Analysis

2.1. Data

Corpus. This study used a corpus of 33 excerpts of singing from 23 songs, with each excerpt having a different tempo. All songs were Mandarin pop songs that were selected to have similar styles. A professional female singer's unaccompanied singing was recorded in a professional recording studio, with a professional sound engineer monitoring the quality of the singing. The recordings were made at a sampling rate of 48 kHz at 32-bit in mono channel. The total length of the songs in this corpus was 111.5 minutes.

Data Annotation The musical beats and the linguistic information were manually annotated by two different annotators. For beat annotation, a music expert with choral conducting experience annotated the beats by adding metronome beats in GarageBand [22] for each excerpt. They noted down the duration from the onset of the audio to the first beat, as well as its tempo (i.e. beats per minute). Two other annotators with experience of playing music instruments listened to the audio files with metronome beats. Disputed parts were returned to the music expert to re-annotate or excluded when no consensus could be achieved. For the linguistic segmentation, a student assistant manually segmented all Mandarin syllables into onsets

and rimes in all 23 songs. The assistant received daily training in phonetic segmentation. They were trained to follow the principles in [23]. An experienced phonetician checked all the practice annotations during the training, and also cross-checked 20% percent of the final corpus annotation to ensure the accuracy and consistency of the segmentation. The musical and linguistic segmentation processes were kept independent from each other to avoid any subjective presumptions from the annotators. The final annotations were integrated into Praat [24] TextGrids. The onsets and rimes that aligned with musical beats were labelled as `beat`, and the ones that did not were labelled as `no beat`. The location of a beat was calculated as a percentage indicating the relative position of the beat within a segment with respect to the total segmental duration. Onsets and rimes are important units in Mandarin and often function as a whole structure (e.g. Tone-Bearing Unit) in organising phonology, speech planning, and speech perception. The onsets and rimes in this study were categorised by the phoneme types of the onsets and the syllable structure of the rimes. The phoneme types included fricative, affricate, nasal, stop, liquid, and glide. The syllable structure for the rimes in Mandarin included V (monophthong), VV (diphthong), VN (Vowel + nasal coda), G (glide) can also appear before these structures and were also counted as a part of the rimes (for more information about Mandarin syllable structure and phoneme types, see [25] [26]).

2.2. Results and Discussions

Presence of Beats. The Mandarin dataset consisted of 6653 Chinese characters, among which 65% were aligned with musical beats. The total number of onset and rimes were 6270 and 6902 respectively. Only 26.3% of onsets (1651) received beats, while almost half of the rimes (3353, 48.6%) aligned with beats. In Figure 1, the types of onsets and rhymes were ranked from low to high by the proportion of beat-aligned onsets and rimes. Among all the onset types, beat-aligned nasals accounted for the smallest proportion, while the fricatives had the largest proportion. The results exhibited a pattern contrary to the sonority hierarchy. Nasals, glides, and liquids are high on the sonority scale, but a smaller proportion of these phoneme types had beat-alignment. The non-sonorant affricates and fricatives, however, had higher chances of aligning with beats. This is against our hypothesis 1 which was based on the sonority hierarchy. However, a possible explanation was that sonority hierarchy was overridden by a more important factor. In this case, phoneme duration was likely to be the more important factor: affricates and fricatives are generally much longer than the other types of the consonants; therefore, there were higher chances for beats to fall onto them. We will examine the effect of phoneme duration on predicting beat presence in Section 4.

Location of Beats. Figure 2 displays the beat location for different phoneme types of the onsets (left) and by different syllable structures for the rimes (right). The onsets showed a strong correlation with the sonority hierarchy in that the sonorant onsets such as glides and liquids had beats aligned to almost the mid point of the onsets. However, the alignment became very late towards the end for the non-sonorant onsets such as fricatives and affricates. For the vowels, alignments across all types were at around the first 25% of the rimes. These results support our hypothesis 2 because the beats did align with the vowel onsets and the transition between the preceding consonants and the vowels; moreover, the data illustrated that the more sonorant the segments were, the more vowel-like they were, and thus the earlier the alignments were.
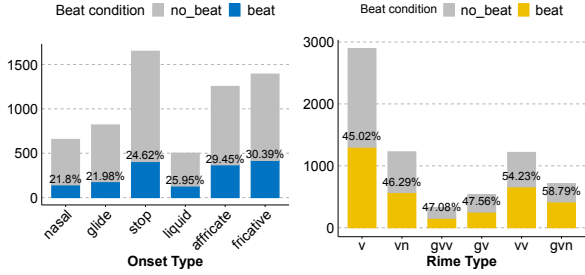
Figure 1: *The numbers and proportions of beat-aligned onsets ([left]) and rimes ([right]) by onset and rime types. Y axis: number of onsets or rimes.*
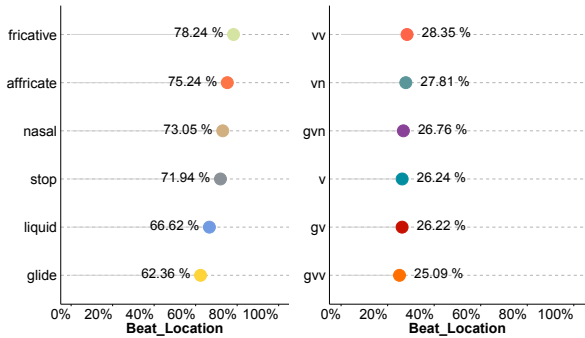

Figure 3: *The numbers and proportions of beat-aligned consonants ([left]) and vowels ([right]) by phoneme types. Y axis: number of consonants or vowels.*

Figure 2: *The average location of beat in relation to the aligned onsets ([left]) and rimes ([right]).*

3. English Singing Analysis

3.1. Data

Corpus. The English singing corpus contained 18 excerpts of singing from 14 English songs, each having a different tempo. The nature of this study required data from professional singers to ensure the beats were correctly produced. However, unaccompanied professional singing data were extremely difficult to acquire. Therefore, we used as many songs as we could access. The data were saved from YouTube as `.wav` files. While the compressed YouTube files may not be suitable for rigorous acoustic analysis, this study only focuses on the timing of the beats in relation to the segments, we therefore considered YouTube audios sufficient for the purpose. The total length of the audio files was 103.4 minutes.

Data Annotation. The same annotators annotated the English singing data following the same procedures. The only difference was the unit of linguistic annotation: in English, consonants and vowels were used instead of onsets and rimes, since onsets and rimes were rarely important units in English. The consonants were categorised into the seven phoneme types according to the characteristics of English: approximant (including glides and liquids), nasal, voiced and voiceless stops, voiced and voiceless fricatives, and affricate. Vowels were divided into monophthongs and diphthongs, which corresponded to V and VV categories in the Mandarin dataset.

3.2. Results and Discussions

Presence of Beats. The English dataset consisted of 3642 English words. 2148 (59.0%) words received musical beats. There
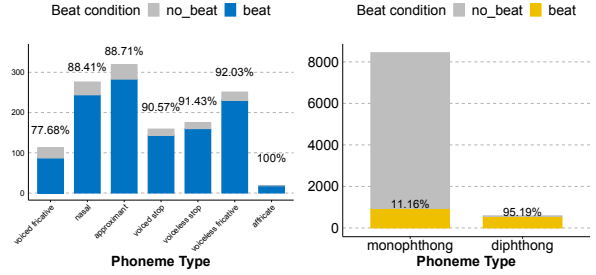
were 10336 segments in this dataset: only 1310 segments were consonants, and the rest of 9026 were vowels. Despite the imbalance between the total numbers of consonants and vowels, the numbers of on-beat consonants and vowels were similar: 1167 consonants and 1496 vowels were aligned with beats. Figure 3 ranks the proportion of beat-aligned consonants and vowels from low to high. Except for voiced fricatives, all other consonant types had very high percentages of segments aligned with beats – more than 88%, although some had smaller absolute numbers. The sonorant phoneme types (nasals and approximants) showed high numbers of beat alignment; however, their percentages were relatively low. The other types did not seem to follow any sonority sequence either. The vowel results showed a higher percentage of diphthongs aligning with beats, which was likely to be related to phoneme length. Based on the results of beat presence from both Mandarin and English data, we can reject the hypothesis that sonorant segments are more likely to receive beats. In contrast with the Mandarin data, which had a much smaller percentage of onsets aligning with beats relative to the rimes, the English results were more unexpected. These results showed a cross-linguistic difference between English and Mandarin data in terms of whether consonants/onsets received a substantial number of beats: around 90% of the consonants were aligned with beats in English, while only approximately 27% of the consonants in Mandarin received beats.

Location of Beats. Figure 4 displays the beat locations for different phoneme types. The consonants show a strong correlation with the sonority hierarchy, with the exception of nasals having relatively late alignment. This may be related to the fact that a larger percentage of nasals were at syllable coda position compared with other phoneme types. The vowel data on the right did not differ much by the vowel types and were aligned at approximately the centre of the vowels. Although sonority did not substantially influence beat presence, it showed an effect on the beat location in both Mandarin and English.

4. Predicting Beats

While we do not aim to provide a predictive model for use in singing synthesis, we fitted several statistical models to predict the presence of beats and the location of beats in order to test the generalisability of our findings. For each task, instead of fitting a single model, we ran the predictive models multiple times with leave-one-out cross-validation. In other words, we reserved data from a song for testing and trained the model with the rest of the data. This process was repeated multiple times until all songs were tested individually, and then the averaged
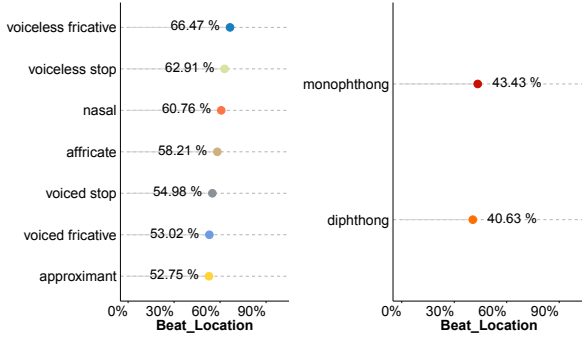
Figure 4: *The average location of beat in relation to the aligned onsets ([left]) and rimes ([right]).*

performance metrics were reported below in Table 1 and 2. Unless otherwise stated, all statistical analyses were perform using the Python package `sklearn` [27].

Predicting the Presence of Beats. Predicting the presence of beats was formulated as a binary classification task using logistic regression with default hyperparameters. Given phonetic features including duration and phoneme types, logistic regression models were trained to predict whether a beat falls on the current phone. The baseline models were simply the majority vote of the majority class (`no beat`). For each language, we fitted three models, one with duration only, one with phone types only and one with both phonetic features included. We did experiment with other features such as syllable structure and word stress but did not find them useful. Therefore they were not reported here. The averaged accuracy for each model over all test sets was presented in Table 1 below. In general, most models outperformed the majority vote baseline, suggesting that there are at least some regularities in terms of the placement of beats. For English, both duration and phoneme type contributed to the prediction, yet phoneme type turned out to be a stronger predictor of beat presence. In contrast, duration was the strongest predictor in the Mandarin model, whereas phoneme type was not more predictive of beat presence than the simple majority vote.

Predicting the Location of Beats. To predict the location of

Table 1: *Predicting beat presence with logistic regression models*

Language	Model	Accuracy
English	Majority vote	0.73
	Duration	0.79
	Phone type	0.89
	Full	**0.92**
Chinese	Majority vote	0.65
	Duration	0.76
	Phone type	0.66
	Full	**0.76**

beats, we framed this as a regression task, in which regression models were employed to predict the exact location of beats within a phoneme as a percentage. For the baseline, the mean value of beat locations in the training set was used as the prediction for all beat locations in the test set. For each language, we fitted two linear regression models using duration and phoneme types respectively, and a full regression with both features. As the relations between beat locations and duration could be non-

linear, we also trained regression trees on all features with a maximum tree depth of five. Averaged mean square errors (MSE) of each model are reported in Table 2. The location of beats was more correlated to phoneme types in both languages, as the duration-only models were not outperforming the mean baseline. Yet in general, even the best performing models in both languages were not highly predictive. The full regression model in Mandarin only brought about 20% decrease of MSE relative to the baseline, whereas the full regression tree model resulted in a 12% decrease compared with the baseline.

Table 2: *Predicting beat location with regression models and regression trees*

Language	Model	Accuracy
English	Mean baseline	0.098
	Regression:Duration	0.111
	Regression:Phone type	0.089
	Regression:Full	0.088
	Regression Tree: Full	**0.086**
Chinese	Mean baseline	0.105
	Regression:Duration	0.129
	Regression:Phone type	0.086
	Regression: Full	**0.082**
	Regression Tree: Full	0.085

5. Conclusions and Future Work

Motivated by the genuine need to resolve bad cases in singing synthesis, this study analysed the synchronisation between speech segments and musical beats in Mandarin and English singing data. Our findings suggest that the P-centre theory of speech rhythm as well as the sonority hierarchy theory can inform segment-beat synchronisation substantially. In summary, the research questions and hypotheses outlined in Section 1.2 can be answered as follows: (1) The sonority of the speech segments did not significantly affect the beat synchronisation. Other factors such as segment duration or syllable structure may have overridden the effect of sonority. (2) The sonority hierarchy and the P-centre theory together predicted the location of the beats: the more sonorant the segments were, the more vowel-like they were, and the earlier the alignments were. (3) Cross-linguistically, the segment-beat alignment in English and Mandarin shared some similarities (universal sonorant phoneme types), but also presented differences (much higher percentage of consonant-beat alignment in English than in Mandarin). In the predictive models, we found that (1) the presence of beats can be predicted by segment duration (English, Mandarin) and phoneme type (English only); (2) the location of beats correlated with phoneme types in both Mandarin and English. These results exhibited cross-linguistic differences, while corroborates with the analysis in the previous sections. Further investigations can be made based on the acoustic correlates of P-centres and sonority hierarchy in order to more accurately predict speech-beat alignment, and contribute to the speech rhythm research.

6. Acknowledgements

7. References

[1] 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 *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2004.11012

[2] Y. Ren, X. Tan, T. Qin, J. Luan, Z. Zhao, and T.-Y. Y. Liu, "Deepsinger: Singing voice synthesis with data mined from the web," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1979–1989.

[3] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer modeling timbre and expression from natural songs," *Applied Sciences*, vol. 7, no. 12, 2017. [Online]. Available: https://www.mdpi.com/2076-3417/7/12/1313

[4] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6189–6193.

[5] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-based singing voice synthesis system," *INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP*, vol. 5, no. December 2014, pp. 2274–2277, 2006.

[6] L. Zhang, C. Yu, H. Lu, C. Weng, C. Zhang, Y. Wu, X. Xie, Z. Li, and D. Yu, "DurIAN-SC: Duration Informed Attention Network based singing voice conversion system," *arXiv*, pp. 1231–1235, 2020.

[7] Y. Hono, S. Murata, K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Recent Development of the DNN-based Singing Voice Synthesis System - Sinsy," *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2018 - Proceedings*, no. November, pp. 1003–1009, 2019.

[8] J. Morton, S. Marcus, and C. Frankish, "Perceptual centers (p-centers)." *Psychological review*, vol. 83, no. 5, p. 405, 1976.

[9] S. M. Marcus, "Acoustic determinants of perceptual center (p-center) location," *Perception & psychophysics*, vol. 30, no. 3, pp. 247–256, 1981.

[10] C. E. Hoequist Jr, "The perceptual center and rhythm categories," *Language and Speech*, vol. 26, no. 4, pp. 367–376, 1983.

[11] A. M. Cooper, D. Whalen, and C. A. Fowler, "The syllable's rhyme affects its p-center as a unit," *Journal of Phonetics*, vol. 16, no. 2, pp. 231–241, 1988.

[12] I. Chow, M. Belyk, V. Tran, and S. Brown, "Syllable synchronization and the p-center in cantonese," *Journal of Phonetics*, vol. 49, pp. 55–66, 2015.

[13] K. de Jong, "Acoustic and articulatory correlates of p-center perception," *UCLA Working Papers in Phonetics*, vol. 81, pp. 66–75, 1992.

[14] A. D. Patel, A. Löfqvist, and W. Naito, "The acoustics and kinematics of regularly timed speech: a database and method for the study of the p-center problem," in *Proceedings of the 14th international congress of phonetic sciences*, vol. 1. Linguistics Department, University of California Berkeley, 1999, pp. 405–408.

[15] P. A. Barbosa, P. Arantes, A. R. Meireles, and J. M. Vieira, "Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[16] T. Rathcke, C.-Y. Lin, S. Falk, and S. D. Bella, "Tapping into linguistic rhythm," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 12, no. 1, p. 11, 2021.

[17] K. L. Pike and C. Fries, *Phonetics: A Critical Analysis of Phonetic Theory and a Technic for the Practical Description of Sounds*, ser. Half-title: University of Michigan publications. Language and literature. Vol. XXI. Bibliography: p. 157-166. University of Michigan Press, 1943. [Online]. Available: https://books.google.nl/books?id=5tFsAAAAIAAJ

[18] G. N. Clements, "The role of the sonority cycle in core syllabification," *Papers in laboratory phonology*, vol. 1, pp. 283–333, 1990.

[19] S. G. Parker, "Quantifying the sonority hierarchy," Ph.D. dissertation, University of Massachusetts at Amherst, 2002.

[20] P. P. K. Mok, "On the syllable-timing of Cantonese and Beijing Mandarin," *Chinese Journal of Phonetics*, vol. 2, pp. 148–154, 2009.

[21] E. Grabe and E. L. Low, "Durational variability in speech and the Rhythm Class Hypothesis," *Laboratory Phonology 7*, pp. 515–546, 2008.

[22] "GarageBand," 2020. [Online]. Available: https://www.apple.com/mac/garageband/

[23] P. Machač and R. Skarnitzl, *Principles of Phonetic Segmentation*. Epocha, 2009.

[24] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2020. [Online]. Available: http://www.praat.org/

[25] S. Duanmu, *The Phonology of Standard Chinese*, 2nd ed. Oxford University Press, 2007.

[26] C. Zhang and X. Wang, "Segment Duration and Proportion in Mandarin Singing," in *Proc. 10th International Conference on Speech Prosody 2020*, 2020, pp. 596–600. [Online]. Available: http://dx.doi.org/10.21437/SpeechProsody.2020-122

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.