Towards Reference-Independent Rhythm Assessment of Solo Singing

Chitralekha Gupta, Jinhu Li and Haizhou Li
Department of Electrical and Computer Engineering, National University of Singapore, Singapore
chitralekha@nus.edu.sg, e0572686@u.nus.edu, haizhou.li@nus.edu.sg

***Abstract*—Rhythm is an important aspect of singing in music information retrieval. From the principles of music theory, note duration is related to the time signature of a song, therefore it provides a representation for rhythm analysis. However, in a reference-independent setup, where the reference musical score of a song is not available, measurement of note duration and hence assessment of rhythm quality of solo singing is difficult. In this work, we analyse the correlation between the duration distributions of musical note lengths and singing syllable lengths, and propose a reference-independent framework that uses singing syllable duration distribution to assess the rhythm quality of solo singing. We show improved prediction of rhythm assessment score using syllable duration histogram compared to existing reference-independent frameworks.**

I. Introduction

Rhythm is an integral aspect of music. Studies have shown that rhythm is one of the most important parameters used by music experts to assess singing quality [1], [2]. Rhythm analysis in music information retrieval is useful for many applications such as song tempo detection, genre classification, and singing quality evaluation [3], [4]. In this study, we analyse the rhythm in solo singing from the perspective of the lyrical content and their correspondence with the musical notes and incorporate a rhythm informed feature to assess the rhythm quality of solo singing.

There have been many studies that implement low-level audio content features to derive beat onsets to characterize the rhythm of a musical piece. They create a representation called a periodicity or beat histogram, which is a measure of the change of the signal amplitude over time [3], [4], [5]. However, these methods rely on the periodic nature of the background music along with the vocals to characterize its rhythm. When background music is absent, i.e. when only the lead melody of the song is present, the note durations (note onsets and offsets) provide a description of the rhythm of the song (this is explained further in Section II). In the case of solo singing when the musical score information is not available, the task of determining note onsets and offsets only from the singing rendition is challenging. Traditionally, the pipeline of note onset and offset detection involved feature extraction, change detection, and a peak detection algorithm [6], [7], [8]. However, unlike pitched instruments where the timbre is usually consistent in the duration of a note, a singing vocal has a higher variance in formant structures within and across notes due to different possible articulation [6]. Pitch irregularities, and pitch modulations also makes the spectrum inconsistent in the duration of a note. Thus, such pipelines for determining note onsets and offsets in singing voices give noisy results.

Typically, to characterize rhythm in solo singing for the purpose of singing quality evaluation, algorithms compared a test singing rendition against the known musical notes of the song [9], [10] or against an ideal singing rendition of the song by a professional singer [11], [12], [2]. These methods extract audio features such as pitch contour and mel-frequency cepstral coefficients (MFCC) that are relevant to perceptual parameters used by music experts. For example, [10], [12] evaluated rhythm consistency by aligning the test pitch contour with the reference pitch contour using Dynamic Time Warping (DTW), and obtained the rhythm score by computing the deviation of the optimal path from a straight line fit in the cost matrix of the DTW between the pitch contours.

Aligning test and reference singing using pitch contour makes rhythm assessment dependent on pitch correctness. So if the test singer sings with inaccurate pitch (off-tune) but maintains a good rhythm, or if the pitch estimation itself is inaccurate, such a method will result in an incorrect assessment of rhythm accuracy. Gupta et al.[11], [2] modified this rhythm assessment measure by using MFCC feature vectors instead of the pitch contour to compute the DTW between reference and test singing. The assumption was that the test singer utters the lyrics correctly, and MFCCs represent the shape of the vocal tract and thus the phonemes uttered. This measure of rhythm assessment was independent of off-tune or incorrect estimates of pitch. However, such reference-dependent methods rely on a reference such as a musical score sheet or an ideal singing rendition, that limits the scope of applications.

Studies have shown that music experts can evaluate singers with a high level of consensus even when the song is unknown to them [13], [14], which suggests that there are inherent shared characteristics of singing quality that differentiate between good and bad singing. This is the motivation for exploring automated methods that evaluate singing quality without depending on a reference singing rendition or music score sheet.

Gupta et al. [15], [16] designed features that characterize

the shape of the pitch histogram and inter-singer distances to evaluate singing quality without a reference. The reference independent methods of singing quality evaluation have mostly involved intonation characterization. In this study, we explore reference-independent rhythm characterization.

We explore and analyse a syllable-informed representation for rhythm assessment of solo singing. We consider that the lyrics of the song are known, and analyse the possibility of using the relation between lyrics and musical notes to derive a rhythm representation of a solo singing rendition. Based on this analysis, we propose a syllable-informed rhythm representation to assess rhythm quality in a solo singing rendition when the musical score information is not available.

II. Music Theory Background for Rhythm Analysis

Rhythm, in music, is often referred to as a strong, regular repeated pattern of sounds, principally according to duration and periodical stress [17], [18], [19]. A *beat* is the basic temporal unit in music, and it is a rhythmic emphasis that happens at regular intervals. *Tempo* is the frequency of beats, and *meter* groups the beats into larger chunks, also at regular intervals. All of these elements together make up the regular repeated pattern of sounds in music, that is rhythm.

In musical notations, rhythm comprises of the *time (or meter) signature*, the *tempo* of the song, and the *relative duration of the notes*. Time signature specifies how many *beats* are contained in each *measure (bar)*, and the value of the basic beat. For example, in Figure 1, the time signature $\frac{2}{4}$ indicates that there are 2 beats per measure, and one beat equals a quarter note (♩). Tempo is the speed at which a piece or a song is meant to be played or sung, measured in beats per minute (bpm). For example, in Figure 1, the tempo is 64 beats per minute, which implies one beat equals 1/64 minutes. According to music theory, the relative duration of notes is a geometric series, i.e. it could be a whole note (𝅝), a half note (𝅗𝅥) (two half notes make a whole note), a quarter note (♩) (two quarter notes make a half note), and so on.Combining all the above definitions together for the example in Figure 1, we get one quarter note equals 1/64 minutes, i.e. 0.9375 seconds.

A. Note Duration Histogram

A comprehensive representation of rhythm in solo singing should encompass all the previously-mentioned aspects of music theory. In earlier studies, beat histogram has been used as the global rhythmic information of a song, but it is primarily useful for detecting the tempo of a song. The main idea behind the calculation of beat histogram is to collect statistics about the amplitude envelope periodicities over multiple frequency bands [4], [20]. The resulting histogram has bins corresponding to tempos in beats per minute (bpm) and the amplitude of each bin corresponds to the strength of repetition of the amplitude envelopes of each channel for that particular tempo - the highest peak corresponding to the tempo of the song. This approach for rhythm representation was derived purely from audio signal analysis, where the periodic structure provided by the background music in the song helped in characterizing the rhythm.

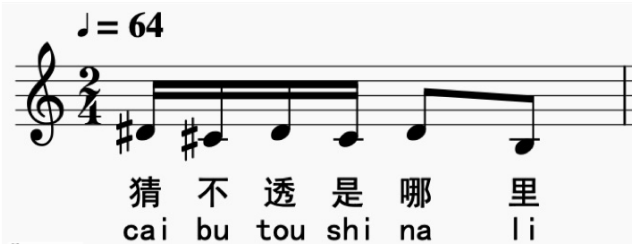


Fig. 1. Excerpt of the musical score sheet of the song 123 木头人.

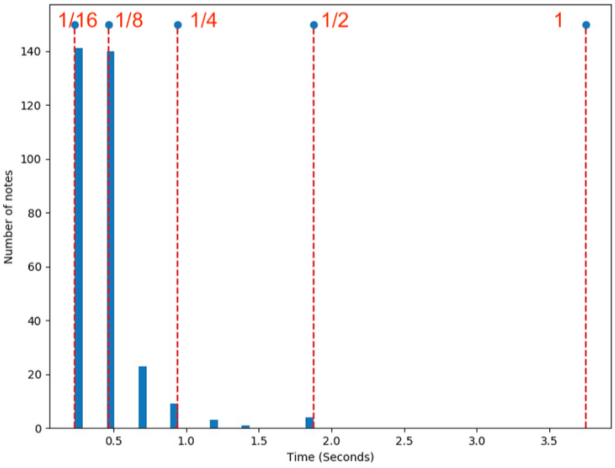


Fig. 2. Histogram of absolute note durations (blue bars) of the song in Figure 1, obtained from the musical notations. The expected durations of whole note, half note, quarter note etc. are shown with red dashed bars.

Previously, the pitch histogram of a singing voice has been shown to provide a comprehensive representation for intonation [15], [16]. But the pitch histogram loses all information about timing, and hence rhythm, and therefore it is not suitable for rhythm analysis. As per music theory, a note duration histogram should provide an adequate rhythm representation of a song. Figure 2 shows the note duration histogram of the song in Figure 1. From the manually annotated digital music score sheet of the song, we computed the length of each note in seconds, i.e. the note durations, and plotted the histogram, where x-axis represents note length in seconds, and y-axis is the number of notes in the song. From our previous discussion, for this song, one quarter note duration equals 0.9375 seconds. Following the geometric progression of relative note duration, eighth note duration will be 0.4688 seconds,

sixteenth note duration will be 0.2344 seconds and so on. From Figure 2, we see that the duration of the frequently occurring notes in the song match with the geometric progression of relative note duration of this song, that was derived from information about the tempo and the time signature of the song. Therefore, the note duration histogram provides a comprehensive representation for rhythm analysis. Next, we present a method to extract the note duration information from singing vocals when digital score sheet of the song is not available.

III. Syllable Duration Histogram as a Rhythm Indicator

The notion of rhythm also occurs in poetry, which is the measured flow of words and phrases as determined by the relation of long and short or stressed and unstressed syllables [21], [22]. In music, prosody is the way the composer sets the lyrics of a vocal composition in the assignment of syllables to notes in the melody.

We demonstrate and confirm the usability of syllable duration as an indicator of rhythm by considering three questions: (1) what is the association between note and syllable, (2) what is the distribution of the difference between syllable duration and note duration, and (3) what is the correlation between syllable duration histogram and note duration histogram.

A. Dataset

Our dataset consists of a total of 100 solo-singing renditions, the average duration of a song is 3.83 minutes and the total duration of the dataset is 6.38 hours. The dataset comprises of 86 unique Chinese pop songs sung by 8 singers (4 male, 4 female), where each singer sings 12 or 13 songs. All the singers are professionally trained in singing with 5-10 years of experience, and all have either performed on stage or have teaching experience. Each singing rendition is manually transcribed in pinyin (which is the official romanization system for standard Mandarin Chinese) and the boundaries of pinyin phones are manually annotated. Additionally, a digital musical score sheet of each singing rendition is manually annotated by music experts in MuseScore[1]. The musical annotations consist of tempo, time signature, note value, and note duration, along with syllable text in pinyin and Chinese characters. This dataset was prepared and verified by Databaker Technologies[2]. Additionally, two native Chinese speaking volunteers listened and verified the phone annotations and boundaries of the dataset.

As the dataset is manually transcribed with pinyin phones, we can derive the syllable boundaries from the phone boundaries with the help of a syllable-to-phone mapping dictionary, and using the Levenshtein distance [23] to align the phones and syllables. We define the syllable's starting phone's start timestamp and syllable's ending phone's end timestamp as the boundaries of the syllable.

[1]https://musescore.com/
[2]https://www.data-baker.com/

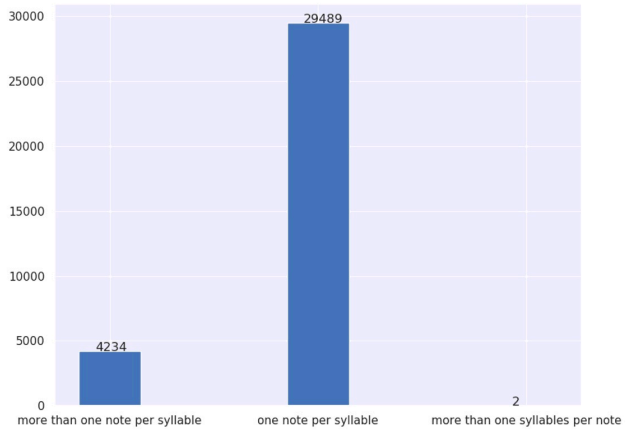


Fig. 3. Distribution of occurrence of different associations between notes and syllables in the dataset.

B. Association between Note and Syllable

From the principles of music composition, one syllable of a word in the lyrics is generally assigned to one musical note [24], [25]. Thus, syllable duration is closely related to the musical note duration, i.e. one steady note duration is likely to correspond to one syllable duration [24]. This relationship between syllable and note duration motivates the use of score information to segment syllables in *Jingju* solo singing renditions by Pons et al.[25]. When more than one note occurs on one syllable, it is called a *melisma*. And can one note have more than one syllable? Figure 3 shows the distribution of occurrences of all of the three cases in our dataset - one note corresponds to one syllable, more than one note corresponds to one syllable, and one note corresponds to more than one syllable. We observe that one note corresponding to one syllable is the most commonly occurring case amongst the three. Therefore, the syllable durations should provide a high correlation with note durations.

C. Distribution of the Difference between Syllable Duration and Note Duration

Figure 4 shows the distribution of the difference between syllable duration and note duration across all syllable/note tokens over all songs. The absolute difference of duration between syllable and note is less than 100 ms for over 81% of the tokens, which confirms that syllable duration is closely related to note duration. Another observation is that 10.2% of the tokens have a longer syllable duration than note duration, whilst 8.5% of the tokens have a longer note duration than syllable duration. Since singing voice has more variability than musical score, therefore we expect to see this slight variation in duration.

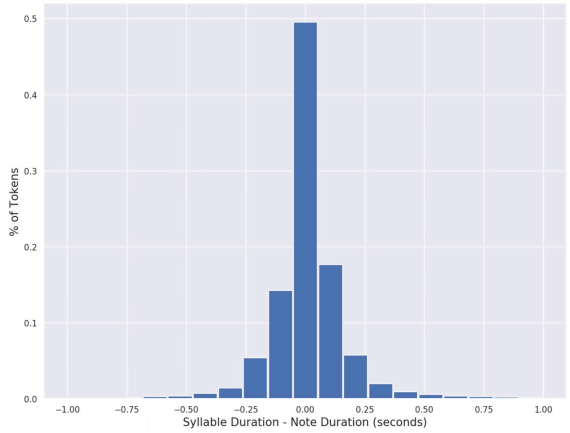


Fig. 4. Distribution of the difference between syllable duration and note duration across all syllable/note tokens over all songs.

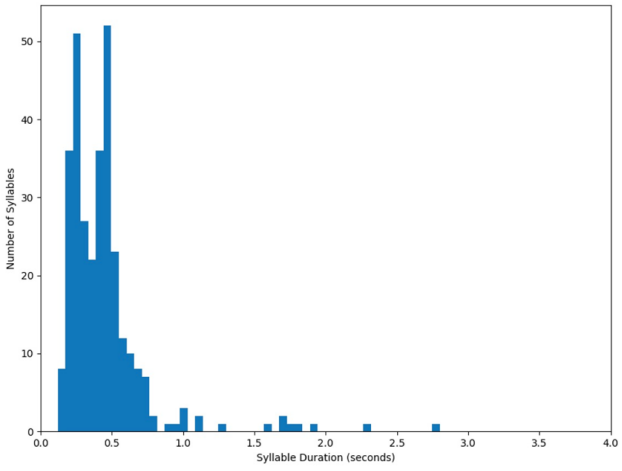


Fig. 5. Histogram of syllable durations of the song in Figure 1.

D. Correlation between Syllable Duration Histogram and Note Duration Histogram

Figure 5 shows the syllable duration histogram of the same song as in Figure 1. Comparing Figure 5 and 2, it is apparent that the two most frequently occurring syllable durations coincide with the two most frequently occurring note durations. Since singing voice has more variability than musical score, therefore we see a spread around the peaks.

To quantitatively verify if the syllable duration histogram correlates with the note duration histogram, we compute the Pearson's correlation coefficient between the note duration histogram and the syllable duration histogram across all the singing renditions in the dataset, where one second equals 10 bins of the histogram. The box plot of this correlation over all the songs of our dataset is given in Figure 6. The average Pearson's correlation between the note duration histogram and the syllable

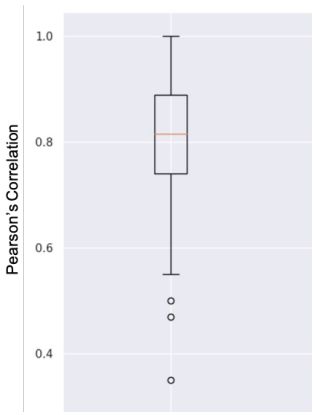


Fig. 6. Distribution of Pearson's correlation between syllable duration histogram and note duration histogram over all songs.

duration histogram across all the 100 songs is 0.80, which shows a strong correlation of the syllable duration histogram with the note duration histogram. This confirms that syllable duration is highly correlated with note duration. This also implies that the pattern of geometric progression of the relative note duration, as described by note duration histogram, can also be reliably represented by syllable duration histogram. Thus, syllable duration histogram provides a reference-independent representation of rhythm in solo-singing.

IV. Rhythm Quality Assessment Framework for Solo-Singing

We propose a lyrics-informed method to derive a rhythm representation of solo singing vocals. We estimate the note duration histogram of singing vocal using syllable durations, and show that they are correlated. We hypothesize that incorporating syllable duration histogram in a solo-singing quality assessment system would help in assessing rhythm quality in a singing rendition.

One thing to be noted is, due to the recent improvement in the performance of the lyrics-to-singing alignment systems [26], [27], we assume that the syllable boundaries can be reliably obtained automatically. Therefore, in this study, we use a dataset where the syllable boundaries were manually marked. In future, an automatic lyrics alignment system will also be a part of this framework.

To systematically examine the use of syllable duration histogram for the purpose of rhythm quality assessment in solo-singing, we prepare an augmented dataset that can be specifically used for rhythm quality assessment in solo singing. We modify an existing automatic singing evaluation framework by conditioning it with syllable duration histogram and train it on the augmented data.

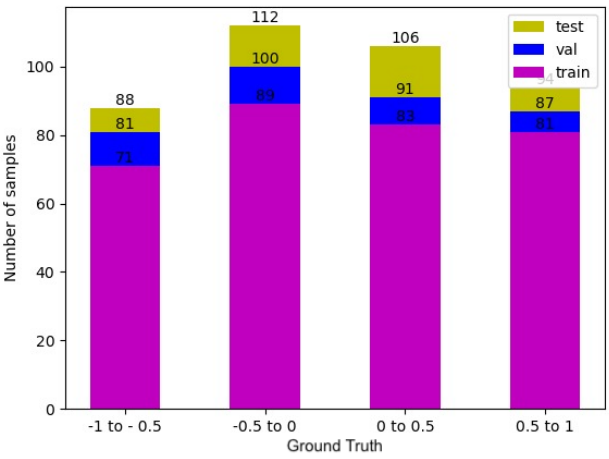
A. Augmented Dataset for Rhythm Quality Assessment

We need a dataset that has human annotations for rhythm assessment for supervised training, however, acquiring a large scale dataset annotated by music experts

for rhythm quality is a difficult and time consuming task. On the other hand, the Databaker dataset (Section III-A) only has recordings from professionally trained singers, and does not have any *negative* samples, i.e. singing vocal examples with incorrect rhythm or varying degrees of rhythm correctness. Therefore, we developed an augmented version of the Databaker dataset to increase the diversity of rhythm quality in the dataset. As rhythm in solo-singing is related to the syllable duration of the words, incorrect duration of syllables would result in incorrect rhythmic structure of the song. In order to synthesize negative rhythm quality samples from the dataset, we generated time-scaled versions of the original singing audio, by modifying the phoneme durations without changing their pitch.

Previously, other works such as Kruspe [28] and Wager et al [29] have applied such time-scaling and other modifications for the purpose of synthesis of examples designed to test particular hypothesis. For example, Kruspe [28] designed a *songified* dataset where a speech dataset was word-wise pitch-shifted and time-scaled to synthesize song-like data for the task of acoustic modeling for singing vocals. Wager et al. [29] synthesized pitch-shifted training examples to train a deep auto-tuner. Inspired by these techniques, we augment the existing professional grade singing vocals dataset with negative singing quality examples where only the rhythm or duration of pinyin phones is modified. It is worthwhile to note here that the purpose of using a dataset that only has duration modifications is to reflect only rhythm quality variations in the dataset, and investigate the usability of syllable-duration histogram for the purpose of rhythm quality evaluation in solo singing, independent of other aspects of singing quality such as intonation. This is the first step towards building a rhythm quality assessment framework for solo singing vocals.

We modify duration of each phone without affecting their pitch using the time-scale modification algorithm based on phase vocoder [30] provided by the library librosa [31]. We vary the duration of each phoneme by a stretch factor r, where $r > 0$ means speeding up the signal, and $r < 0$ means slowing down the signal. Each original singing rendition will have four other duration modified versions with r values of +/- 0.2, 0.4, 0.6, and 0.8. Therefore, there is a total of 400 augmented audio files, in addition to the original 100 audio files. We apply a Gaussian distribution with a variance of 0.05 around the r value such that, for example, the 0.2 version is actually N(+/-0.2,0.05), which will effectively render each phoneme at or around 1.2 or 0.8 times its original speed. The ground-truth rhythm quality assessment score is assigned between -1 to 1 where 1 is set to the original singing vocal, and for the rest the ground-truth score is set to $gt = -2 \times |r| + 1$. So when $r = 0.2$, $gt = 0.6$. This augmented dataset is divided into train, validation, and test sets at a ratio of 8:1:1. We call the test set in this augmented dataset as *test set 1*. Figure 7 shows the distribution of samples across the different ranges of ground-truth values.



(a)

Fig. 7. Distribution of augmented data across different ranges of ground-truth values in train, validation, and test sets.

B. Test dataset with human annotations

Additionally, in order to test our proposed framework on actual singing, we make use of the publicly available dataset[3] presented in [11], that we call PESnQ_Dataset. This test dataset consists of 20 solo singing renditions, each from a different singer, along with professional music expert assessment of each of these singing renditions based on various musically relevant perceptual parameters such as pitch (intonation accuracy), and rhythm (rhythm consistency), each assessed separately on a likert scale of 1 to 5. These human annotated ground-truths are mapped to a range of -1 to 1, to make it similar to *test set 1*. We will refer to this test dataset as *test set 2*.

For obtaining the syllable boundaries, we applied the solo-singing acoustic model presented in [32]. A factorized time-delay neural network (TDNN-F) model was trained in Kaldi [33] using solo singing data with MFCC and i-vectors, as described in detail in [32]. The mean and median absolute word boundary error of this model reported on solo-singing test data was 200 ms and 30 ms respectively. We forced-align the lyrics (split into syllables) to the solo-singing renditions of the PESnQ_Dataset using this solo-singing acoustic model, and obtained syllable boundaries automatically.

C. Rhythm Quality Assessment Framework

Recently, Huang et al. [34] studied a singing quality evaluation framework, which was based on a framework derived from the work on music performance assessment by Pati et al [35]. We re-purpose these frameworks for rhythm quality assessment by training them with our augmented dataset. We train three versions of this framework:

[3]https://github.com/chitralekha18/PESnQ_APSIPA2017

one, which is the baseline convolutional recurrent neural network (CRNN) with Constant-Q transform (CQT) as the input representation [35], [34]. The second is the hybrid-CRNN which is the baseline CRNN conditioned on the pitch histogram (PH-CRNN) [34]. We modify this framework by replacing the pitch histogram with the syllable-duration histogram, i.e. we condition the CRNN framework by concatenating its intermediate representation with the syllable duration histogram vector (SH-CRNN). The proposed SH-CRNN framework is shown in Figure 9. We compare the performance of the baseline CRNN, PH-CRNN, and SH-CRNN for the task of rhythm quality assessment in terms of Pearson's correlation between the predicted score and the ground-truth score *gt* for test sets 1 and 2.

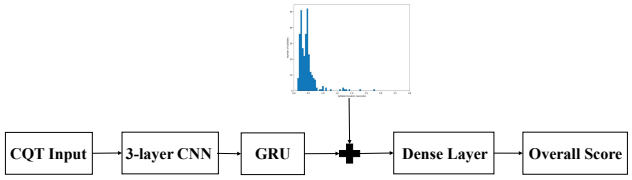
1) Input Representation: We use Constant-Q transform or CQT as the input 2D representation of the audio. CQT uses geometrically spaced frequency bins to ensure that the the ratio (Q) of the center frequencies to bandwidths of all bins are constant. Huang et al [34] showed that CQT performs better than other representations such as Mel spectrogram and Chromagram for singing quality evaluation. CQT is specifically seen to be well suited for music data, since the Q factor is approximately constant in most of the audible frequency range of the human perception system, and the fundamental frequencies of the tones in Western music are geometrically spaced along the standard 12-tone scale. For CQT computation, the window length and hop size are set to 2,048 and 512 respectively. There are 96 bins per CQT and 24 bins per octave to capture sharp/flat pitches. The calculated CQT is squared and then scaled into decibels (dB).

We compute the syllable duration histogram using the syllable boundaries derived from the manually marked phone boundaries as discussed in Section III-A. The histogram is computed such that one second equals 10 bins of the histogram, with a total of 30 bins, and we normalize this histogram such that the area under the histogram integrates to 1 (i.e. density).

2) Network Architecture: The CRNN network consists of 3-layer CNNs. Each CNN sub-structure contains a 2D convolutional layer, a 2D batch normalization layer, an exponential linear unit (ELU) activation function and a 2D max-pooling layer. Then it is followed by 1-layer RNN with gated recurrent units (GRUs). Cost function is calculated as the mean squared error between model prediction and the ground-truth. We used the Adaptive Moment Estimation optimizer and trained our model for 250 epochs. For each epoch, the batch size is equal to 5 for training, validation and test sets.

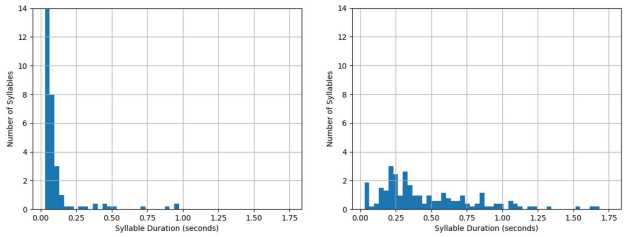
D. Results

1) Qualitative Observation: Figure 9 shows the syllable histograms of two singing renditions from *test set 2* - one from a good singer with a human rating of 4.5 out of 5.0 for rhythm quality and the other from an amateur singer rated



(a)

Fig. 8. Proposed automated rhythm quality evaluation framework SH-CRNN, i.e. CRNN conditioned with syllable duration histogram.



(a) (b)

Fig. 9. Syllable Duration Histogram for singing rendition of the song *I have a dream (ABBA)* with (a) good rhythm (rated 4.5/5.0 by music experts), and (b) poor rhythm (rate 1.0/5.0 by music experts).

1.0 out of 5.0. The syllable histograms clearly show that the good singer has clear peaks indicating that the syllables are consistent in duration, whereas the amateur singer shows a dispersed syllable histogram, indicating inconsistency in syllable durations. This observation supports our hypothesis that syllable duration histogram provides a representation of rhythm that is a useful indicator for assessing rhythm correctness in singing.

2) Quantitative results: Table I shows that the framework conditioned on the syllable duration histogram improves the performance of rhythm quality assessment score prediction both on validation and *test set 1*, compared to the baseline CRNN system. This means that explicitly encoding rhythm accuracy related information via syllable duration histogram assists in reliably predicting rhythm quality assessment score.

The last column in Table I shows that SH-CRNN outperforms the baseline systems and shows a positive correlation with human annotations of rhythm quality for *test set 2* that consists of actual singing. The correlation values are lower for *test set 2* than *test set 1*. This indicates that although duration modified data introduces variability in rhythm quality in the training data, there could be artefacts introduced because of the time-scaling algorithm that may negatively impact the model. Moreover, artificially modified singing data does not completely represent the real examples of rhythm variations in actual singing vocal. Therefore, further investigation is required to adapt the model with a dataset containing actual singing examples with different rhythm qualities.

TABLE I
COMPARISON OF THE PERFORMANCE OF OUR PROPOSED HYBRID CRNN CONDITIONED ON SYLLABLE DURATION HISTOGRAM (SH-CRNN) WITH THE BASELINE CRNN FRAMEWORK AND PH-CRNN FRAMEWORK, ON TEST SETS 1 AND 2.

Framework	Model description	Train	Val	Test 1	Test 2
CRNN[35], [34]	The CRNN model using CQT as input	0.97	0.59	0.70	-0.42
PH-CRNN [34]	The hybrid CRNN model using CQT and pitch histogram as inputs	0.99	0.58	0.74	-0.04
SH-CRNN	The hybrid CRNN model using CQT and syllable duration histogram as inputs	0.87	**0.81**	**0.75**	**0.32**

V. CONCLUSIONS

In this work, we propose a musically-motivated representation for rhythm quality assessment of solo-singing renditions. According to music theory, a note duration histogram provides a comprehensive representation for rhythm analysis. We show that syllable duration is closely related to note duration, and we quantitatively verify that a histogram of syllable durations has a strong correlation with the note duration histogram. Moreover, we applied the syllable duration histogram for the task of rhythm accuracy assessment in singing vocals without a reference, and showed improved prediction of rhythm assessment score using syllable duration histogram compared to a baseline system. This is the first step towards reference-independent rhythm quality assessment in solo-singing vocals. Our code-base is available online[4]. In future, a lyrics alignment system will be integrated with this framework to obtain syllable boundaries automatically. Further investigation of this methodology needs to be done on a solo-singing test dataset that includes naturally variable rhythm quality in singing (as opposed to the artificially designed dataset in this work).

This study opens up opportunities for automatic rhythm analysis by leveraging the link between note duration and syllable duration. This direction of research can take advantage of the note detection algorithms that are linked to note duration analysis, and the speech recognition algorithms that are linked to syllable duration analysis. This study can be applied for other tasks such as tempo detection or a comprehensive and explainable assessment of singing quality.

ACKNOWLEDGMENT

This research work is supported by Academic Research Council, Ministry of Education (ARC, MOE), Singapore. Grant: MOE2018-T2-2-127. Title: Learning Generative and Parameterized Interactive Sequence Models with RNNs.

REFERENCES

[1] C. Cao, M. Li, J. Liu, and Y. Yan, "A study on singing performance evaluation criteria for untrained singers," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on*. IEEE, 2008, pp. 1475–1478.

[2] C. Gupta, H. Li, and Y. Wang, "A technical framework for automatic perceptual evaluation of singing quality," *APSIPA Transactions on Signal and Information Processing*, vol. 7, 2018.

[3] A. Lykartsis and A. Lerch, "Beat histogram features for rhythm-based musical genre classification using multiple novelty functions," *10.14279/depositonce-9530*, 2015.

[4] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[5] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *Proceedings of the AES 25th International Conference*, 2004, pp. 196–204.

[6] C.-C. Toh, B. Zhang, and Y. Wang, "Multiple-feature fusion based onset detection for solo singing voice." in *ISMIR*, 2008, pp. 515–520.

[7] S. Chang and K. Lee, "A pairwise approach to simultaneous onset/offset detection for singing voice using correntropy," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 629–633.

[8] H. Heo, D. Sung, and K. Lee, "Note onset detection based on harmonic cepstrum regularity," in *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2013, pp. 1–6.

[9] W.-H. Tsai and H.-C. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1233–1243, 2011.

[10] E. Molina, I. Barbancho, E. Gómez, A. M. Barbancho, and L. J. Tardón, "Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 744–748.

[11] C. Gupta, H. Li, and Y. Wang, "Perceptual evaluation of singing quality," in *Proceedings of APSIPA Annual Summit and Conference*, vol. 2017, 2017, pp. 12–15.

[12] C.-H. Lin, Y.-S. Lee, M.-Y. Chen, and J.-C. Wang, "Automatic singing evaluating system based on acoustic features and rhythm," in *Orange Technologies (ICOT), 2014 IEEE International Conference on*. IEEE, 2014, pp. 165–168.

[13] T. Nakano, M. Goto, and Y. Hiraga, "Subjective evaluation of common singing skills using the rank ordering method," in *Ninth International Conference on Music Perception and Cognition*. Citeseer, 2006.

[14] ——, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Ninth International Conference on Spoken Language Processing*, 2006.

[15] C. Gupta, H. Li, and Y. Wang, "Automatic leaderboard: Evaluation of singing quality without a standard reference," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 13–26, 2020.

[16] ——, "Automatic evaluation of singing quality without a reference," in *Proceedings of APSIPA Annual Summit and Conference*, 2018.

[17] P. Fraisse, "Rhythm and tempo," *The psychology of music*, vol. 1, pp. 149–180, 1982.

[18] G. T. Toussaint, *The geometry of musical rhythm: what makes a" good" rhythm good?* CRC Press, 2019.

[19] N. Stojkoski, *Essay on Human Reason: On the Principle of Identity and Difference*. Vernon Press, 2018.

[20] G. Tzanetakis, G. Essl, and P. Cook, "Human perception and computer extraction of musical beat strength," in *Proc. DAFx*, vol. 2, 2002.

[21] D. W. Harding, D. C. Harding, and D. W. Harding, *Words into rhythm: English speech rhythm in verse and prose*. Cambridge University Press, 1976.

[22] G. Fant, A. Kruckenberg, and L. Nord, "Stress patterns and rhythm in the reading of prose and poetry with analogies to music performance," in *Music, language, speech and brain*. Springer, 1991, pp. 380–407.

[4] https://github.com/AME430/TOWARDS-REFERENCE-INDEPENDENT-RHYTHM-ASSESSMENT-OF-SOLO-SINGING.git

[23] G. Navarro, "A guided tour to approximate string matching," *ACM computing surveys (CSUR)*, vol. 33, no. 1, pp. 31–88, 2001.
[24] E. Nichols, D. Morris, S. Basu, and C. Raphael, "Relationships between lyrics and melody in popular music," 2016.
[25] J. Pons Puig, R. Gong, and X. Serra, "Score-informed syllable segmentation for a cappella singing voice with convolutional neural networks," in *18th International Society for Music Information Retrieval Conference; Suzhou, China*. International Society for Music Information Retrieval (ISMIR), 2017.
[26] C. Gupta, E. Yılmaz, and H. Li, "Automatic lyrics alignment and transcription in polyphonic music: Does background music help?" in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 496–500.
[27] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 181–185.
[28] A. M. Kruspe and I. Fraunhofer, "Training phoneme models for singing with" songified" speech data." in *ISMIR*, 2015, pp. 336–342.
[29] S. Wager, G. Tzanetakis, S. Sullivan, C.-i. Wang, J. Shimmin, M. Kim, and P. Cook, "Intonation: A dataset of quality vocal performances refined by spectral clustering on pitch congruence," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 476–480.
[30] J. Driedger and M. Müller, "A review of time-scale modification of music signals," *Applied Sciences*, vol. 6, no. 2, p. 57, 2016.
[31] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.
[32] C. Gupta, E. Yılmaz, and H. Li, "Acoustic Modeling for Automatic Lyrics-to-Audio Alignment," in *Proc. Interspeech 2019*, 2019, pp. 2040–2044. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1520
[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.
[34] L. Huang, C. Gupta, and H. Li, "Spectral features and pitch histogram for automatic singing quality evaluation with crnn," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 492–499.
[35] K. A. Pati, S. Gururani, and A. Lerch, "Assessment of student music performances using deep neural networks," *Applied Sciences*, vol. 8, no. 4, p. 507, 2018.