Training Explainable Singing Quality Assessment Network with Augmented Data

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

***Abstract*—Data-driven methods for automatic singing quality assessment have so far focused on obtaining an overall singing assessment score of a given singing rendition. However, the *explainability* of such a score in terms of musically relevant components of singing quality such as intonation accuracy and rhythm correctness has not been attempted due to the lack of annotated training data. In this work, we propose to augment a singing vocals dataset, containing only professional singing renditions, with negative samples for improving the diversity in singing quality examples in the training data. We validate this augmented dataset through listening tests. Moreover, we use this data to formulate a multi-task learning framework that can simultaneously provide pitch accuracy feedback along with an overall singing quality score for a given singing rendition. We show that our methods outperform existing systems for both unseen songs and singers singing English and Mandarin popular songs.**

I. Introduction

Singing is a popular art form and is a desirable skill for people to learn. Daily vocal singing practice as well as professional assessment scenarios depend on professional music experts or musicians to evaluate the performance of singers. But for amateurs, feedback from music teachers is not always conveniently available. Moreover some professional assessment scenarios, such as music grade examinations, and singing competitions can benefit from a singing assessment system which can provide objective assessment and feedback while saving costs. Therefore, there is a need to build an automated and reliable singing skill evaluation system that could be useful for singing pedagogy, singing contests, and karaoke systems, in turn making singing training more accessible to singing enthusiasts. Ideally, such a system should provide an interpretable feedback to the performer, for example how well have they performed on the different musical parameters such as intonation accuracy, and rhythm consistency.

Many earlier studies have shown that intonation accuracy and rhythm consistency are the perceptual parameters that music experts rely on the most in their evaluation process [1], [2]. Automated singing quality evaluation studies have designed algorithms that can objectively define and assess these perceptual parameters. These studies can be broadly divided into two approaches – reference dependent [1]–[6], where a test singing rendition is compared against a standard reference singing rendition or musical score, and reference independent [7]–[12], where the assessment algorithm depends on properties of singing quality and supervision from music experts. Although reference dependent approaches allow one to define explainable objective measures, they make the method dependent on the choice of a reference sample. Nakano et al. [7] argued and showed that music experts do not rely on a reference singing audio to assess a singing rendition of unknown melodies, rather relied on the inherent characteristics of the singing quality.

In recent works [10], [11], [13], neural network frameworks have been employed to learn implicit features from the time-frequency representations of the singing vocal audio allowing the model to learn the inherent characteristics of singing voice through supervised learning while not depending on a reference singing rendition. Although these methods have achieved good performance, they depend on a dataset annotated by music experts, which is not easily scalable. Moreover, they are trained to only give an overall assessment score, as such a score is the only human annotation practically available for large datasets. Therefore, such systems fail to provide detailed feedback to the singers about their singing quality in terms of musical parameters such as pitch accuracy, and rhythm correctness. In this work, we design a method to construct an augmented dataset for the task of singing quality assessment. Furthermore, we propose an evaluation framework that provides pitch accuracy feedback along with an overall singing quality evaluation score. We believe our proposed approach is the first step towards building a scalable and explainable singing quality assessment system. Our contributions are:

1) A method to develop an augmented dataset containing negative samples from the original professional singing vocals. The augmented data samples are validated through listening tests.
2) A multi-task learning framework for explainable singing quality assessment that evaluates the overall performance as well as pitch correctness.

II. Related Works

In the field of music information retrieval, publicly available singing vocal datasets are scarce for singing quality evaluation and the existing public datasets often have some inherent defects such as poor audio quality, biased distribution of songs and incorrect labels [14]. There has been some effort in building singing vocals datasets for the purpose of singing quality assessment [4], [9], however the number of audio recordings is small, and manual annotations are only available for overall singing quality. Most of the public datasets provided are dependent data collection method and audio labeling by human annotators [15], [16] from their laboratory or company, which is time-consuming and expensive.

Singing vocals dataset such as NHSS [17] and NUS48E [18] only have professional singing vocals recordings. Such datasets are not balanced for the purpose of singing quality assessment because amateur singing qualities are under-represented. A subset of the DAMP dataset [9], consisting of 400 singing renditions, was assessed by human annotators for their overall singing quality through a crowd-sourcing platform. However, for training an explainable neural network model for singing quality assessment, apart from overall singing quality score, annotations for the individual perceptual parameters such as intonation accuracy, rhythm consistency etc. are also needed. The lack of appropriate datasets restricts research in building a holistic singing quality evaluation system.

In many recent studies, multi-task learning has shown improvement in performance of the main task by the learning from related tasks [19], [20]. In MIR, some recent works [21], [22] have used a common feature extraction structure and allowed the network to learn multiple related tasks, that has achieved better evaluation results. For example, Zhang et al. [23], [24] applied multi-task learning for emotion detection, and multi-task frameworks are also verified on singing voice separation by Stoller et al. [25]. However, to the best of our knowledge, there is no research to build a singing evaluation system based on multi-task learning and provide multiple feedback.

III. Data Augmentation

In order to build a data-driven automatic singing quality evaluation framework, datasets with balanced distribution of singing performance audio examples of different levels of singing abilities along with their assessment scores are essential. Kruspe [26] 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, as real singing vocals with words boundaries was not available at the time. Wager et al. [27] proposed a deep autotuner or pitch correction network where they synthesized their training examples by a de-tuning process where they pitch-shifted every note by up to 100 cents (1 semitone). Inspired by these techniques, we augment existing professional grade singing vocals datasets with negative singing quality examples. This data augmentation step is important to diversify the singing dataset in a controlled manner so that singing quality assessment frameworks can be trained. In this section, we discuss our data augmentation method in detail. The subjective and objective validation experiments are discussed in Section V.

A. Pitch-shifting and Speech samples

As discussed in Section II, some publicly available singing vocals datasets such as NHSS and NUS48E only have professional singing vocals recordings. For the purpose of increasing singing quality diversity in such datasets, we generate pitch-shifted versions of the original singing audio.

Setting the original professional singing rendition as the best example, we generate negative examples by inducing pitch deviations in the original. We change the pitch values based on word boundaries. These word boundaries are either provided in these singing vocals datasets or can be automatically obtained by audio-to-lyrics alignment algorithms that have seen success in recent times[1].

We generate de-tuned pitch-shifted samples, i.e. negative samples, by randomly raising or lowering the pitch of each word segment extracted from the original professional singing rendition along the continuous logarithmic scale of cents. For each word, a value can be randomly selected from a list of user-defined pitch offset values and either added or subtracted from the original pitch values of the word. By constructing the list of pitch offset values, many different de-tuned negative examples can be generated. In the study of intonation deviation using midi by Wager et al. [27], they chose pitch deviations that were less than or equal to 200 cents (i.e. two semitones), in order to focus their analysis on intonation behavior when the singer deviates from the expected pitch, yet is close to it, while noting that larger differences could be due to other reasons such as misalignment of notes in time, noisy extraction process etc. Thus, in our study, we choose pitch offset values within 200 cents from the original pitch and use Pitch Synchronous Overlap Add (PSOLA) algorithm provided in librosa library for pitch-shifting.

In addition to the pitch-shifting examples, we also desire to have a sample that can be used as the worst-performing sample in our augmented datasets to contrast with the trained singer's renditions. It can be safely assumed that if one sings in a manner of speaking, which is monotonous in terms of pitch, then it can be considered to be the worst kind of singing performance. Since datasets such as NHSS and NUS48E have read-lyrics version of each song, we use those as the worst singing samples. However to match with the singing style, we modify the speech sample such that the duration of each word is same as the duration of the corresponding singing word. We modify duration without affecting the pitch using the time scale modification algorithm based on phase vocoder [28] provided by the library librosa. On rare occasions, when a word in the speech recording is not present at the corresponding location in the singing vocal recording, we skip modifying that word.

B. Pitch Score Ground Truth

One of the drawbacks of existing singing quality evaluation datasets [9] is that they only have an overall assessment score and do not have detailed manual annotation about perceptual parameters such as intonation accuracy, rhythm consistency etc. Therefore, supervised training for such detailed parameters is not possible with the existing datasets. With our augmented

[1] https://www.music-ir.org/mirex/wiki/2020:Automatic_Lyrics-to-Audio_Alignment_Results

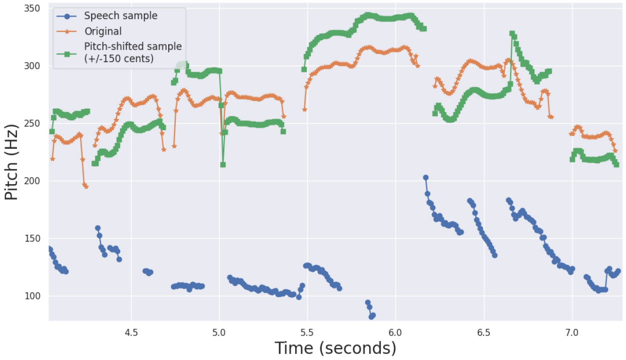


Fig. 1. Comparison of three kinds of pitch score ground-truths.

dataset, we can provide a pitch accuracy indicator in proportion to the amount of pitch deviation introduced. However, to the best of our knowledge, there is no theory to suggest that a linear continuous variation in the pitch deviation would correlate with a linearly varying perceptual score. In this work, we only provide three classes as pitch score ground truth for the supervised learning framework - the original audio renditions as ‘good’, the pitch-shifted synthesized renditions as ‘medium’ and speech samples as ‘bad’ because of their obvious difference in pitch correctness. This is the first step towards providing an explainable singing assessment feedback. In future, further investigation and human studies are required to understand the relation between pitch deviation and perceptual score to provide a detailed pitch accuracy score .

Figure 1 illustrates the pitch contours of three versions of an audio sample from the augmented dataset. The pitch of each phoneme/word of the pitch offset sample will deviate from the original pitch value by a fixed value. For example, the 150-cent pitch-shifting rendition is generated by raising or lowering the pitch of each word by 150 cents. For the modified speech sample, the overall pitch is very low and flat and sound like an out-of-tune rendition.

IV. Explainable Framework

Our goal is to design a framework that provides useful feedback to assist users in learning singing technique and performance improvement. The feedback should be more than an overall assessment score that can inform the singer about the various perceptual parameters used by music experts to assess singing quality, such as pitch accuracy, rhythm consistency, etc. [1]. In this study, we design a multi-task neural network framework to provide pitch accuracy feedback along with the overall score.

We build upon the framework presented by Lin et al. [10] that predicts only the overall score, but with pitch histogram embedding appended to the embedding extracted from the spectral representation. The pitch histogram, which is the distribution of pitch values in a singing rendition [29], has been previously shown to be a strong descriptor of singing quality [9]. The audio snippet is chosen as the input of the neural network. We obtain the spectro-temporal representation (Mel-spectrogram, Constant-Q Transform (CQT) and Chromagram) of this audio and feed it into the Convolutional Recurrent Neural Network (CRNN) to get the assessment of this performance. In the pitch histogram of a good singing rendition, there are several narrow, sharp, and well-defined spikes that indicate that the dominant notes are hit repeatedly and consistently, while a poor quality singing rendition has a dispersed distribution of pitch values, that reflect that the singer is unable to hit the dominant notes of the song consistently. We believe that the pitch histogram would be a supportive indicator for pitch quality assessment, while a combination of pitch assessment representation and spectro-temporal representation of the singing rendition would capture the overall singing quality.

We explored three multi-task frameworks for predicting pitch assessment score along with overall assessment score as shown in Figure 2.

In *Framework 1*, we use two separate branches with independent weights, one for pitch evaluation and the other for overall evaluation. In pitch evaluation branch, we combine the pitch histogram as an embedding feature with the features from the CRNN network. The weighted sum of errors from the two branches is used to backpropagate and updates the weights. In *Framework 2*, we merge the features extracted independently from the two CRNN networks with the pitch histogram, and use a fully-dense network to obtain a two-dimensional vector output through the full-connected layer. The first dimension of this vector is pitch score prediction and the other is overall score prediction. In *Framework 3*, we combine the embedding from the pitch score branch with the extracted embedding from the CRNN to predict overall score.

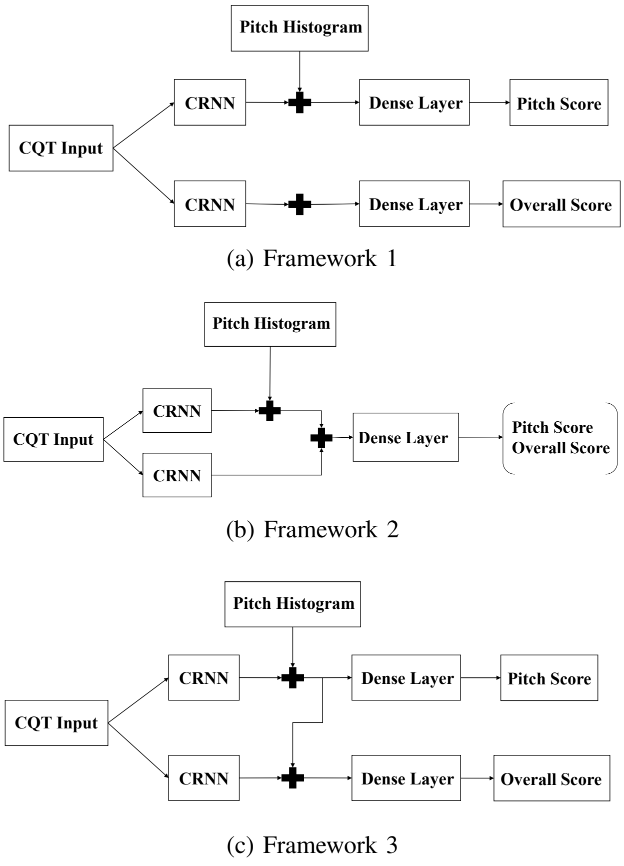
Studies previously have shown that perceptually, the overall score is a combination of various perceptual parameters [1], [2]. Although in this study we have only one of the perceptual parameters, i.e. pitch accuracy, the motivation of exploring these three kinds of frameworks was to understand the inter-dependence between the pitch score and overall score branches.

V. Experiments and Results

A. Datasets

1) DAMP subset: DAMP[2] is a singing vocals dataset provided by the company Smule[3]. The singing renditions provided in DAMP were recorded through Smule’s Sing! karaoke mobile app by singing enthusiasts around the world, but with limited control over the recording conditions. A subset of this dataset was curated by [9] and was later also used by [10]. There are four popular English songs in this DAMP subset. For each song, there are 100 different performers with mixed levels of singing skills, no common performers between songs, and equal number of males and females. Each singing audio is divided into 5 segments, each segment is 20-30 seconds long

[2] https://ccrma.stanford.edu/damp/
[3] https://smule.com



(a) Framework 1

(b) Framework 2

(c) Framework 3

Fig. 2. Multi-task Frameworks

and full-length audio sample is also available. The dataset is divided into train, validation, and test sets at a ratio of 8:1:1, as in [10].

In this dataset, each performer's singing is also provided with a Best-Worst Scaling (BWS) score which was obtained from pairwise comparison between singing vocals from volunteers with music expertise on a crowd sourcing platform, Amazon mechanical turk by [9]. This score gives a rank-ordering of the singers in each song. BWS scores was used in these studies on account of the fact that people are better at relative judgments, like choosing a preferred singing rendition between two given samples, rather than giving an exact score [30]. The BWS score is defined as follows:

$$B = \frac{n_{best} - n_{worst}}{n} \tag{1}$$

where n_{best} and n_{worst} are the number of times a singer is marked as preferable and otherwise, and n is the total number of times the singer appears in the pairwise BWS tests. BWS score is regarded as the ground-truth in the dataset and its value range is between -1 and 1. We obtained this dataset and its annotations from the authors upon request.

2) Databaker: This is a singing vocal recordings dataset of Mandarin pop songs purchased from the company Databaker[4]. It consists of a total of 101 singing audio files (86 unique songs) sung by 8 singers (4M/4F) recorded in a sound-proof professional recording studio. Each singer sings 12 or 13 songs. All the singers are professionally trained in singing for 5-10 years, and all have either performed on stage or have teaching experience. This dataset also consists of the read version of the lyrics (spoken lyrics) recorded by the same speakers. Each singing and spoken recording is manually transcribed in pinyin and the boundaries of pinyin phones are manually marked.

However, this singing vocals dataset is not balanced for training a singing quality evaluation framework, since all audio samples in this dataset are recorded from trained singers with good singing quality. So in Section V-C, we applied our data augmentation technique to this dataset to generate negative samples. The augmented dataset is divided into train, validation, and test sets at a ratio of 8:1:1 and the proportions of the original samples, pitch-shifting samples and speech samples are very close among train, validation, and test sets.

3) NHSS: Sharma et al. [17] present a database of parallel recordings of speech and singing called NUS-HLT Speak-Sing (NHSS) dataset, which is available on request. This dataset consists of 102 singing recordings of English pop songs in total from 10 professional singers (5 female and 5 male). Each singer sings about 10 songs which are selected from a list of 20 songs. The English word boundaries are provided both for speech and singing renditions. In terms of sample distribution, this dataset is similar to the Databaker dataset so the same data augmentation method is applied to this dataset. The augmented dataset is divided into train, validation, and test sets at a ratio of 8:1:1 and the proportions of the original samples, pitch-shifting samples and speech samples are very close among train, validation, and test sets.

4) NUS-48E subset: NUS Sung and Spoken Lyrics Corpus(NUS-48E corpus) was presented by Duan et al. [18]. A subset of this dataset was annotated for singing quality by music experts [4]. This subset is the only singing performance dataset with comprehensive perceptual scores such as overall score, pitch score, and rhythm score. There are two songs and each song is performed by 12 singers, so there are 24 audio samples in total. Although this dataset is well-calibrated, the amount of recordings is too small to train a neural network, so it will be used to fine-tune the pitch score prediction model.

In Figure 3 and Table 1, we provide statistics of the original and augmented versions of these datasets.

All the datasets consist of pop genre of songs, even though they are sung in different languages. Since the features being used and evaluated in this work are related to the prosody of the song, and not the language, therefore we have assumed language-independence in this work.

[4] https://test.data-baker.com

B. Evaluation Metric

We use Pearson's correlation coefficient [31] to measure linear correlation between the singing quality evaluation score predicted by our model and the BWS score in the datasets that was annotated by humans, and use this as an indicator of performance of our models. Higher the correlation value, better is the singing quality evaluation prediction of the model, as it will be closer to how humans would judge.

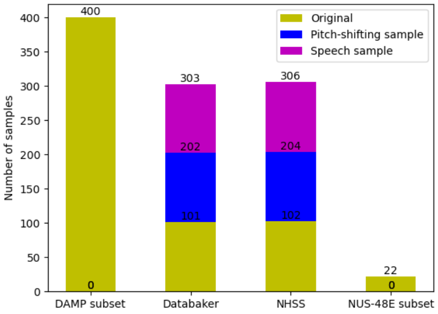


Fig. 3. Number of Samples for Datasets

C. Validation of Data Augmentation

We validate our data augmentation method through subjective listening tests and objective verification experiment by showing that the augmented dataset improve the reliability and performance of the current singing quality evaluation system.

1) Subjective evaluation experiment: We conducted listening tests on a selected subset of the augmented DataBaker and NHSS datasets. We used four versions of audio samples: the original professional singer's rendition, 75-cent pitch-shifting rendition, 150-cent pitch-shifting rendition and speech sample.We chose these four kinds of samples to verify whether the new samples are significantly different from the original samples in terms of pitch. We choose different artificial pitch offsets (75 cents and 150 cents in this subjective evaluation experiment) to observe the influence of pitch offsets on the generated samples. At the same time, we compared the effect with pitch-shifting samples and speech samples. Twelve different songs were used out of which half were Chinese songs (from Databaker) and half were English songs (from NHSS) to avoid language disparity. For each audio file, four audio snippets each of 15-20 seconds were selected, so the total number of singing snippets were 48. These constraints were imposed in order to limit the evaluation time for each volunteer within 15 minutes and to ensure the quality of their judgement. The listeners were asked to rank order the four versions presented to them based on their judgment of their pitch correctness with regardless of subtle difference in sound quality.

TABLE I
Dataset Statistics

Dataset	# Original Audio Recordings	Language	Gender Distribution
DAMP subset	400	English	200(M) 200(F)
Databaker	101	Chinese	4(M) 4(F)
NHSS	102	English	5(M) 5(F)
NUS-48E	22	English	5(M) 6(F)

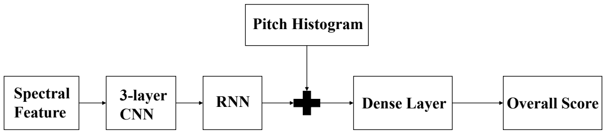


Fig. 4. Framework in Objective Verification Experiment.

We invited sixteen student volunteers who are not professional musicians to participate in this listening test and statistically analyzed the results. From the listening test results, we observe that inter-judge correlation amongst the 16 volunteers is 0.83, which means the judges mostly agree with each other. Such an inter-judge correlation was also observed in other studies [4]. We expect that an audio sample that has higher pitch shifting value will rank lower. In that sense, the original professional singing recording should be rank 1, the 75 cents pitch shifted version will be rank 2, the 150 cents version will be rank 3, and the speech sample will be rank 4. The Pearson's correlation between the average rank-order provided by the volunteers and our expected rank-order is high, at 0.96. This result supports our hypothesis that the data augmentation method helps in diversifying the singing skills samples by adding negative samples to the datasets consisting only of trained singing renditions. It is worth noting that even if the volunteers we invited were students rather than professional musicians, they still easily and highly consistent gave the rank of these four samples in terms of pitch correctness, which confirmed the big difference between the generated samples and the original samples.

2) Augmented data for training singing quality assessment framework: In order to validate that the augmented datasets can assist in the task of singing quality evaluation and improve its performance, we used the augmented datasets to re-train the Hybrid-CRNN singing evaluation framework from Huang et al. [10] and compared the performance.

The framework in [10] which is illustrated in Figure 4 consists of a pitch histogram embedding appended to the spectro-temporal embedding extracted from a CRNN. The authors trained and evaluated this framework on the train and test sets of the DAMP-subset and assessed the performance of the framework with three spectro-temporal input representations - Mel-spectrogram, Constant-Q Transform and Chromagram. In our study, we pre-trained the same framework on the Databaker and NHSS augmented datasets and then utilized this model to train and evaluate on the DAMP subset with Pearson's correlation coefficient. We keep the parameters for spectro-temporal input representations and the structure and hyper-parameters of the network consistent with [10] in the whole process. Table II compares our results with that from [10] by calculating the correlation between the predictions by the neural network model and the ground-truth in the dataset. The results show an improvement in performance with the pre-trained augmented dataset model for the better performing features CQT and Chromagram.

Additionally we wanted to test whether our model trained with augmented data can provide reliable assessment results on unseen songs and singers which means some songs and singers that do not appear in the training and validation sets. Therefore, we conducted leave-one-song out (LOSO) and leave-one-singer out (LOSI) experiments by selecting one or some of the songs or singers (about 10% of the total samples) as the test set and won't be trained by our model, the remaining songs or singers are used as the train and validation set to train our model. These results are based on the evaluation metric of Pearson's Correlation Coefficient and averaged over cross-folds on the DAMP subset dataset with the same Hybrid-CRNN framework with CQT input features. Table III shows that our pre-trained augmented data model provides a significant improvement over the available result from [10].

Through these experimental results, we conclude that our data augmentation method provides an automatic way of creating negative samples in existing trained singer recordings datasets that diversifies these datasets to re-purpose them for the task of singing quality evaluation.

D. Explainable Framework

Our goal is to provide an overall singing quality evaluation along with a pitch correctness feedback. One challenge is that we do not have a dataset with both overall score and pitch evaluation score to train the multi-task model. So we designed an additional *pseudo pitch score ground-truth* prediction model to provide pseudo pitch score labels for the DAMP subset. Then we use this DAMP subset with extended annotations to train our multi-task learning frameworks. In this section, we discuss the pseudo pitch score ground truth prediction model, analyse the three proposed multi-task frameworks, and present a comparative study of our best performing results with other existing systems.

TABLE II
PEARSON'S CORRELATION COEFFICIENT RESULT WITH HYBRID-CRNN FRAMEWORK UNDER DIFFERENT INPUT SPECTRAL FEATURES

	Mel	CQT	Chromagram
Huang [10]	0.63	0.76	0.74
Ours	0.50	0.78	0.75

TABLE III
UNSEEN SONGS AND SINGERS EXPERIMENT WITH HYBRID CRNN. (LOSO: LEAVE-ONE-SONG-OUT; LOSI: LEAVE-ONE-SINGER-OUT)

	LOSO	LOSI
Huang [10]	0.56	NA
Ours	0.61	0.59

1) Pseudo pitch score ground truth: We train the pitch score classification framework in Figure 5 with the augmented Databaker and NHSS datasets along with their 3 pitch score class ground-truths as explained in Section III-B. The ground-truth is represented as a 3-dimensional one-hot vector that correspond to the three pitch score classes. We also fine-tune the model on the small NUS-48E corpus. This pseudo pitch score prediction model showed a high Pearson correlation of 0.98 for the test sets of Databaker and NHSS datasets.

Using this model, we predicted the *pseudo pitch score ground truth* of all the audio samples in DAMP-subset. The prediction is a three-dimensional output vector and the dimension with the maximum value is regarded as the classification label for that audio sample.

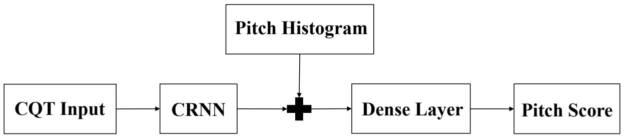


Fig. 5. Pseudo Pitch-score Ground Truth Prediction Model

2) Training of the Networks: For CQT input, the hop size is set to 512 and we set 96 bins per CQT and 24 bins per octave to capture sharp/flat pitches. 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. For our multi-task framework, cost function is calculated as the weighted sum of mean squared error for overall and mean squared error for pitch score with weights of 0.7 and 0.3 respectively, where the weights are chosen empirically. 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.

3) Evaluating the Proposed Multi-task frameworks: The three multi-task frameworks presented in Section IV were trained and tested on the DAMP subset. The Pearson's correlation between the predicted overall score of the test dataset from these frameworks and the corresponding human-annotated BWS scores is shown in Table IV. The third framework, that consists of a combination of the features learnt from the input representation and the pitch score branch, provides the best performance amongst the three frameworks. In framework 1, the pitch score and overall score prediction branches function independently except for the combined error that is back-propagated. Therefore, there is not much learning between the branches. In framework 2, the pitch score and overall score influence each other through the common dense layer. However, framework 3 most closely corresponds to the interpretation of overall score by music experts, i.e. overall score is a combination of the evaluation of individual perceptual parameters [1]. The overall score, in this framework, learns

TABLE IV
COMPARISON OF MULTI-TASK FRAMEWORKS

	Pearson Correlation
Framework 1	0.74
Framework 2	0.75
Framework 3	0.77

TABLE V
FRAMEWORK 3 EVALUATION (S.=SCORE; C.=PEARSON'S CORRELATION; CLASS.ACC.=CLASSIFICATION ACCURACY)

	Overall S.C.	Pitch S.C.	Pitch S. class. acc.
Train	0.99	0.96	96%
Validation	0.77	0.90	94%
Test	0.77	0.91	94%

from a combination of input representation embedding and the pitch score branch embedding.

Table V shows multi-task framework 3 performance on the DAMP subset for overall singing quality and pitch correctness score. The pitch score classification accuracy of this model is 94% on the test set along with a 0.77 Pearson correlation result for continuous overall score. In our dataset, the ground truth of the pitch score is stored as a three-dimensional vector. The correlation between the three-dimensional vector predicted by our explainable framework and the pitch score label is 0.91 on the test set which indicates that our framework can reliably classify audio samples according to intonation performance.

4) Comparison with existing work: We compared our framework with the CPH-CRNN model which is best in [10] and other existing architectures [9], [12] on the same DAMP subset test set, as shown in Table VI. It is clear that our model improves the overall evaluation result through the auxiliary learning of the pitch evaluation task. Further more, our model can provide a reliable pitch evaluation feedback, which is beneficial for users to improve their performance.

VI. CONCLUSIONS

We proposed a method to augment existing singing vocal datasets with negative samples, validate the method through listening tests, and show the use of the augmented dataset for the task of automated singing quality assessment. This method is a simple and controlled way to scale up and diversify the training dataset when there is a lack of examples of all kinds in the training set or it is difficult to obtain music expert evaluation labels. Additionally, we also propose a multi-task singing quality evaluation framework, that makes use of the controlled augmented dataset to provide a more precise overall score along with a pitch score feedback. Our code-base is available online[5]. Our proposed framework is the first step towards a comprehensive explainable framework for singing quality assessment that can help singing enthusiasts to hone their skills.

TABLE VI
COMPARISON WITH EXISTING WORK, IN TERMS OF PEARSON CORRELATION OF THE PREDICTED OVERALL SCORES WITH HUMAN BWS SCORES, AND THE PREDICTED PITCH SCORES WITH PSEUDO PITCH GROUND TRUTH SCORES.

Framework	Overall Score	Pitch Score
Ours	0.77	0.91
Huang et al [10]	0.76	NA
Pati et al [12]	0.56	NA
Gupta et al [9]	0.48	NA

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 *2008 9th International Conference on Signal Processing*. 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] 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.

[4] C. Gupta, H. Li, and Y. Wang, "Perceptual evaluation of singing quality," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 577–586.

[5] 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 *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 744–748.

[6] J. Huang, Y.-N. Hung, A. Pati, S. K. Gururani, and A. Lerch, "Score-informed networks for music performance assessment," *arXiv preprint arXiv:2008.00203*, 2020.

[7] 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.

[8] N. Zhang, T. Jiang, F. Deng, and Y. Li, "Automatic singing evaluation without reference melody using bi-dense neural network," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 466–470.

[9] 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, 2019.

[5] https://github.com/AME430/Towards-Training-Explainable-Singing-Quality-Assessment-Network-with-Augmented-Data.git

[10] 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.

[11] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset." in *ISMIR*, 2018, pp. 468–474.

[12] 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.

[13] C. Gupta, L. Huang, and H. Li, "Automatic rank-ordering of singing vocals with twin-neural network."

[14] C. Gupta, R. Tong, H. Li, and Y. Wang, "Semi-supervised lyrics and solo-singing alignment." in *ISMIR*, 2018, pp. 600–607.

[15] B. Bozkurt, O. Baysal, and D. Yüret, "A dataset and baseline system for singing voice assessment," in *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR), Matosinhos, Portugal*, 2017, pp. 25–28.

[16] R. Gong, R. C. Repetto, and X. Serra, "Creating an a cappella singing audio dataset for automatic jingju singing evaluation research," in *Proceedings of the 4th International Workshop on Digital Libraries for Musicology*, 2017, pp. 37–40.

[17] B. Sharma, X. Gao, K. Vijayan, X. Tian, and H. Li, "Nhss: A speech and singing parallel database," *arXiv preprint arXiv:2012.00337*, 2020.

[18] 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 *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2013, pp. 1–9.

[19] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[20] M. Yang, W. Zhao, W. Xu, Y. Feng, Z. Zhao, X. Chen, and K. Lei, "Multitask learning for cross-domain image captioning," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1047–1061, 2018.

[21] S. Böck, M. E. Davies, and P. Knees, "Multi-task learning of tempo and beat: Learning one to improve the other." in *ISMIR*, 2019, pp. 486–493.

[22] Y.-N. Hung and A. Lerch, "Multitask learning for instrument activation aware music source separation," *arXiv preprint arXiv:2008.00616*, 2020.

[23] B. Zhang, G. Essl, and E. M. Provost, "Recognizing emotion from singing and speaking using shared models," in *2015 international conference on affective computing and intelligent interaction (acii)*. IEEE, 2015, pp. 139–145.

[24] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5805–5809.

[25] D. Stoller, S. Ewert, and S. Dixon, "Jointly detecting and separating singing voice: A multi-task approach," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 329–339.

[26] A. M. Kruspe and I. Fraunhofer, "Training phoneme models for singing with" songified" speech data." in *ISMIR*, 2015, pp. 336–342.

[27] 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.

[28] J. Driedger and M. Müller, "A review of time-scale modification of music signals," *Applied Sciences*, vol. 6, no. 2, p. 57, 2016.

[29] G. Tzanetakis, A. Ermolinskyi, and P. Cook, "Pitch histograms in audio and symbolic music information retrieval," *Journal of New Music Research*, vol. 32, no. 2, pp. 143–152, 2003.

[30] J. J. Louviere, T. N. Flynn, and A. A. J. Marley, *Best-worst scaling: Theory, methods and applications*. Cambridge University Press, 2015.

[31] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.