AUTOMATIC RANK-ORDERING OF SINGING VOCALS WITH TWIN-NEURAL NETWORK

Chitralekha Gupta[1] **Lin Huang**[1] **Haizhou Li**[1,2]

[1] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2] Machine Listening Lab, University of Bremen, Germany

chitralekha@nus.edu.sg, lin.huang@u.nus.edu, haizhou.li@nus.edu.sg

ABSTRACT

When making judgements, humans are known to be better at choosing a preferred option amongst a small number of options, rather than giving an absolute ranking of all the options. This preference-based judgment rank-ordering method is called Best-Worst Scaling (BWS). Inspired by this concept, we propose a preference-based framework to generate a relative rank-ordering of singing vocals, and therefore, singers. We adopt a twin-neural network (Siamese) that learns to choose a preferred candidate in terms of singing quality between two inputs. With a few such pairwise comparisons, this method generates a relative rank-order of a complete list of singers. Additionally, we incorporate a knowledge-based musically-relevant pitch histogram representation, as a conditioning vector, to provide explicit musical information to the network. The experiments show that this method is able to reliably evaluate singing quality and rank-order singing vocals, independent of the song or the singer. The results suggest that the twin-neural network learns the underlying discerning properties relevant to singing quality, instead of being specific to the content of a song or singer.

1. INTRODUCTION

Singing is a popular form of entertainment and a desirable skill to develop [1]. In recent times, many online applications that provide a platform to showcase singing talent as well as socially engage through music have become popular, such as Smule Sing! [1], Starmaker [2], Quanmin K Ge [3], and SoundCloud [4]. With high volumes of singing performances on such online platforms, there is a need to explore automated methods of assessing the quality of singing for the purpose of identifying singing talent as well as providing meaningful feedback to amateur and aspiring singers. For example, such an automated evaluation method would be useful for screening of the singers for popular singing talent reality shows such as American Idol and The Voice. In this work, we provide a data-driven and preference-based framework for evaluating singing quality.

Previous work on automatic singing quality evaluation has focused on comparing a test singing rendition against the known musical notes of the song [2, 3] or against an ideal singing rendition of the song by a professional singer [4–6]. These methods extract audio features such as pitch contour and mel-frequency cepstral coefficients that are relevant to perceptual parameters used by music experts to evaluate singing quality such as intonation accuracy, rhythm consistency, and timbre brightness [7,8]. However, such methods are constrained by the need for a reference or ideal singing rendition for every song. Moreover, the choice of an "ideal" reference singer introduces a bias of subjective choice.

Another approach is the assessment of singing quality without a reference singer. Studies have shown that music experts can evaluate singers with a high level of consensus even when the song is unknown to them [9, 10], which implies that there are underlying inherent characteristics of singing quality that differentiate between preferred and amateur singing. Previously, Nakano et al. [10] designed features such as pitch interval accuracy that measure the offset of the pitch contour from the musical semitone grid to evaluate singing quality without a reference rendition. Gupta et al. [11, 12] designed hand-crafted features that characterize the shape of the pitch histogram and inter-singer distances to evaluate singing quality without a reference. Such methods provide insight and explanation to the objective evaluation, such as the measurement of the sharpness of peaks in a pitch histogram correlating with the consistency of hitting musical notes. However, such hand-crafted features provide an approximate representation of singing quality, that depend on manual thresholds that are determined empirically. They do not capture all aspects of singing and therefore are limited.

Previously in [11], the authors showed that since a song can be sung correctly in one or a few similar ways, but incorrectly in many different and dissimilar ways, it implies that the quality of a singer is proportional to his/her similarity with other singers. However, to obtain a relative rank-order based on this idea, they needed to calculate the

[1] https://www.smule.com/
[2] https://www.starmakerstudios.com/
[3] https://kg.qq.com/
[4] https://soundcloud.com/

distance of every singer in the dataset with respect to every other singer, which becomes computationally demanding as the size of the dataset increases. Moreover, this distance calculation made sense only if the singers were singing the same song, making the algorithm song-dependent.

Humans are known to be better at relative judgments, i.e. choosing the most preferred singer among a small set of singers, than giving an absolute rating [13, 14]. This is the basis of the best-worst scaling (BWS) method used for consumer value preference surveys [15]. Motivated by this human behavior, we would like to develop a singing evaluation framework that is song-independent. The task is to rank-order a list of singing vocals without the need of any singing reference. We achieve a rank-ordering of a long list of candidates through a number of pairwise decisions.

2. TWIN-NETWORK FOR RELATIVE SINGING QUALITY EVALUATION

Twin-neural networks (or Siamese networks) have been previously used to measure similarity between two audio inputs, for example for vocal imitation [16–18], singing style identification [19, 20], and singing query retrieval [21]. The idea behind using twin-neural network for the task of singer style identification is to map different singing and song renditions of the same singer closer to each other than those of different singers. However, to the best of our knowledge, twin-neural network has not be explored for the task of singing quality assessment.

In this work, we modify a twin-neural network such that it learns which of the two given singing inputs is *more preferable* in terms of singing quality. We then obtain the rank-ordering of singing vocals by counting the number of times a singing input is preferred in many such pairwise comparisons across different singers, based on the concept of BWS. A similar approach has been discussed by Niu et al. [22] where a twin-neural network is applied for the task of image quality assessment. The network learns to rank the quality scores between the two input image patches, where it applies cross entropy as the loss function. Our work differs from [22] in that we propose a novel and intuitive preference metric and comparative loss function for training a siamese neural network to predict ranking.

Additionally, we include explicit musical knowledge in this framework, by using the pitch histogram as a conditioning vector. The two arms of the twin-network share the same architecture as well as parameters, i.e., the two inputs pass through exactly the same networks for feature learning. Singers share a similar underlying singing vocal production mechanism, however they differ in quality due to prosodic characteristics such as the ability to consistently hit the right notes. We hypothesize that the two arms of the network should be able to project each singing vocal to a compressed latent space that only represents the discriminatory singing quality properties independent of the song or the singer, thus making it suitable for the task of singing quality comparison of two singing vocals. Furthermore, BWS rank-ordering method is known to provide a reliable rank-ordering with fewer number of comparisons,

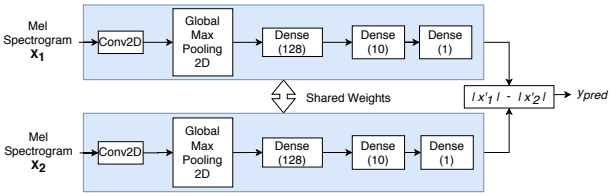


Figure 1. Twin-neural network modified for preference-based singing quality judgment.

which is helpful when the dataset size increases.

2.1 Model

A twin neural network consists of two identical sub-networks that have the same configuration with the same parameters and shared weights. During training, the parameter update is mirrored across both the sub-networks. The two sub-networks extract features from the two inputs, and then similarity between the two feature vectors is computed by a distance metric. In general, in a twin neural network, for a pair of inputs (x_1, x_2), the distance metric of the output of the two sub-networks $f(x_1)$ and $f(x_2)$ is given by Euclidean distance

$$D = ||f(x_1) - f(x_2)||_2 \tag{1}$$

The contrastive loss function, that needs to be minimized, is defined as

$$L = y_t \cdot max(1 - D, 0)^2 + (1 - y_t) \cdot D^2 \tag{2}$$

where y_t is the ground truth label, such that $y_t = 1$ whenever x_1 and x_2 are from the same class and $y_t = 0$ otherwise. This framework has been successfully used for similarity detection tasks such as sound search and vocal imitation detection [16, 18].

We modify this framework such that it learns to choose the better singer amongst the two input singers, as shown in Figure 1. To do this, we propose to replace (1) the distance metric with a *preference metric*, and (2) the contrastive loss with a *comparative loss*.

2.2 Preference Metric

We define the preference metric as the difference between the L1 norm of the feature vectors,

$$D_p = |f(x_1)| - |f(x_2)| \tag{3}$$

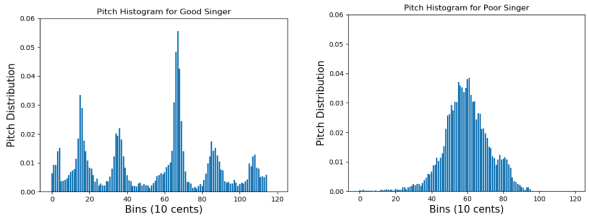
where $|f(.)|$ is the L1 norm of the feature vector. This provides a direction to the comparison, i.e. if $D_p \geq 0$ implies singer 1 input rendition x_1 is better than or similar to singer 2 input rendition x_2, and $D_p < 0$ implies singer 2 is better than singer 1. In contrast, a distance metric can only provide the magnitude, but not the direction of the difference.

2.3 Comparative Loss

Given the preference metric D_p, we compute the comparative loss function to be minimized, as

$$L_c = y_t \cdot max(1 - D_p, 0)^2 + (1 - y_t) \cdot (D_p + 1)^2 \tag{4}$$

where, y_t is the ground truth label, such that $y_t = 1$ whenever x_1 is better than or similar to x_2, and $y_t = 0$ otherwise. Note that the modification in comparative loss compared to contrastive loss is to accommodate for the directional or signed property of the preference metric D_p. Let's



(a) (b)

Figure 2. Pitch histograms of (a) a preferred singer (rank 1) and (b) an amateur singer (rank 99) from the song *Let it go* of dataset 1 (Section 4.1.1). (1 bin = 10 cents).

examine this equation closely. If x_1 is better than x_2, then $y_t = 1$, so equation 4 will become

$$L_c = max(1 - D_p, 0)^2 \tag{5}$$

Minimizing this loss function, makes the preference metric D_p close to 1. On the other hand, if x_2 is better than x_1, then $y_t = 0$, so equation 4 will become

$$L_c = (D_p + 1)^2 \tag{6}$$

For this loss function to be zero, the preference metric D_p should be optimized to -1, thus preserving the signed property of the preference metric D_p.

3. HYBRID TWIN-NEURAL NETWORK WITH PITCH HISTOGRAM CONDITIONING VECTOR

We use mel-spectrogram as the input time-frequency representation of the input audio waveforms. However, measuring pitch correctness is a vital component of singing quality evaluation. Therefore, we condition the twin-network with pitch information in the form of pitch histogram. This unburdens the network from learning pitch-related information from the input representation.

The pitch histogram represents the distribution of pitch values in a sung rendition [23]. As demonstrated by [11], a pitch histogram is a strong indicator of the quality of singing. A pitch histogram is computed as the count of the pitch values (calculated in the unit of cents) folded on to the 12 semitones in an octave, where one semitone represents 100 cents on equi-tempered octave. The melody of a song typically consists of a set of dominant musical notes (or pitch values). These are the notes that are hit frequently in the song and sometimes are sustained for long duration. In the pitch histogram of a preferred singing rendition, there are several narrow, sharp, and well-defined spikes that indicate that the dominant notes are hit repeatedly and consistently (Figure 2(a)). On the other hand, an amateur 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 (Figure 2(b)).

Due to its strong relevance to singing quality, we condition the twin-neural network by concatenating the pitch histogram vectors of the two inputs, ph_A and ph_B to the output vector of their respective sub-network intermediate layer, as shown in Figure 3. Such a configuration, called the hybrid twin-neural network, improves the performance

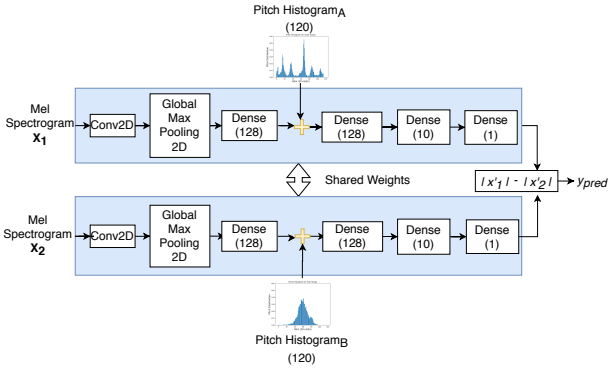


Figure 3. Hybrid Twin-neural network conditioned on pitch histogram.

of the network for singing quality evaluation, by providing explicit pitch-related information to the network releasing degrees of freedom to the network to learn other non-pitch related properties.

4. EXPERIMENTAL SETUP

We conduct experiments to evaluate the performance of the twin-neural network and the hybrid twin-neural network for the task of automatic rank-ordering of singing vocals. We analyse the performance of the framework when the input pair of singing vocals belong to the same song, as well as when they belong to different songs. We also compare the performance and capabilities of this framework with previous similar work in literature.

4.1 Dataset

4.1.1 Singing voice dataset 1

We use the subset of DAMP dataset [5] that was curated by [11] for the purpose of singing quality evaluation. It consists of solo-singing recordings (16 kHz sampling rate, mono) of 4 popular Western songs each sung by 100 unique singers (50 male, 50 female). There were no common singers across different songs. The selection of songs was based on the available number of unique singers in the DAMP dataset, and equal distribution between males and females, to avoid gender bias. The 4 popular songs are *Let it go (Idina Menzel), Cups (Anna Kendrick), When I Your Man (Bruno Mars), Stay (Rihanna)*. All the songs are rich in steady notes and rhythm, as summarized in Table III of [11].

We use one 20-30 seconds long snippet from each singing rendition. This snippet is a common section of the song for all the singers singing that song. The ground-truth subjective ranking provided with this dataset was a BWS score obtained through a crowd-sourcing platform by asking listeners to choose the best and the worst amongst a few singers singing the same song. This score resulted in a rank-order of the singers of each song from 1 to 100, where rank 1 means the best singer, and rank 100 means the worst singer. We divide this dataset into a train set that has 80%, i.e. 80 singers per song, and validation and test sets, each

[5] https://ccrma.stanford.edu/damp/

Dataset	Division	#songs	#singers per song	#singer pairs per same song	#singer pairs singing same+different songs
1	Train	4	80	4 x 80 x 79 = 25,280	80x4=320 singers 320x319=102,080 pairs
	Validation	4	10	4 x10 x 9 / 2 = 180	10x4=40 singers 40 x 39 / 2 = 780 pairs
	Test	4	10	4 x 10 x 9 / 2 = 180	10x4=40 singers 40 x 39 / 2 = 780 pairs
2	Test	2	10	2 x 10 x 9 / 2 = 90	10x2=20 singers 20 x 19 / 2 = 190 pairs

Table 1. Summary of the number of singer pairs from the different datasets used in this work.

consisting of 10%, i.e. 10 singers per song. To ensure similar distribution of singing quality in all of these subsets, we pick the singers with ranks [1,11,21,...91] for the test set, [2,12,22,...92] for the validation set, and the rest for the train set. Note that the test set consists of singers that are not present in training or validation sets. However, the songs in all three sets are the same.

4.1.2 Singing voice dataset 2

To test the trained models on unseen songs, we use a small test dataset provided by [4] that consists of solo-singing recordings (16 kHz sampling rate, mono) of 2 Western pop songs (*I have a dream (ABBA), Edelweiss (Sound of Music)*) each sung by 10 singers. Since this dataset was recorded in a lab-controlled environment, the entire spectrum of singing ability - from amateur singers to professionally trained excellent singers - was covered. The ground-truth singing quality annotations provided in this dataset are absolute ratings on a scale of 1-5, 5 being the best, provided by professional music teachers and/or performers in a lab-controlled environment. Additionally, the music experts evaluate and score the quality of pitch and rhythm separately.

4.1.3 Singer pair inputs

The number of singer pairs singing the same song and different songs in the two datasets is summarized in Table 1. There exist a total of 25,280 ordered pairs of singers singing the same song in the training set, and 102,080 ordered pairs of singers singing different as well as same song. We treat ordered pairs, i.e. singer pairs (A,B) and (B,A) as different training samples for the purpose of data augmentation. Also this is helpful because the preference metric is asymmetric. The validation set consists of 180 unordered pairs of singers singing the same, 780 unordered pairs of singers singing the same and different songs. We have two test sets, one each from singing vocals datasets 1 and 2. The number of pairs in test dataset 1 is same as the validation set. The test dataset 2 consists of 90 unordered pairs of singers singing the same song, and 190 unordered pairs of singers singing the same and different songs.

4.1.4 Ground-truth Labels

The ground-truth label for each pairwise comparison of singers is derived from the human scores provided in the singing datasets 1 and 2. In dataset 1, the BWS score b provided for every singer that ranges between -1 and 1 is first normalized between 0 and 1. We then label a pair of singers A and B as

$$y_{true} = \begin{cases} 1, & \text{if } b_A \geq b_B \\ 0, & \text{otherwise} \end{cases}$$

which implies that if singer A is better or similar to singer B, the label is 1, and it is 0 otherwise. Similarly, in dataset 2, the absolute human ratings between 1 and 5 is normalized between 0 and 1, and the same method is applied to give a binary label every pair of singers.

4.2 Setup

The overall structure of the twin network is shown in Figure 1, and the hybrid twin network is a modified version of the twin network, as shown in Figure 3.

4.2.1 Pre-processing

Both the input audio waveforms are converted to the 2-D mel-spectrogram representations with 96 mel-bins over the frequency range of 0–8 kHz, a window length of 512 and a hop-size of 256. The input waveforms are singing vocal snippets of approximately 20 seconds in duration. The number of frames of the two inputs are made equal by appending zeros to the shorter spectrogram.

4.2.2 Twin-network

Each arm of the twin network consists of a 2D Convolutional Neural Network (CNN) with 1 convolutional layer having 64 filters with a kernel size of 3x3 and stride size of 1x1, followed by a sigmoid activation function. They are then each followed by a 2D global max-pooling layer, and three fully-connected dense layers. There are 128 neurons in the first dense layer, 10 in the second layer, and 1 in the third. The sigmoid activation function is used in all of these layers, as it squashes the output of the layers between 0 and 1. Empirically, we observe that applying the sigmoid activation at all the layers results in convergence while training.

The preference metric is computed using the outputs of these two arms, as discussed in Section 2.2. This value is viewed as the preference judgment value between the input singer pair, i.e. which of the two singers is better.

4.2.3 Hybrid twin-network

In order to incorporate musical relevance into the network, we concatenate the normalized 120 dimensional pitch histogram vectors of the two inputs, at the output of the first dense layer in both the arms of the twin network. We chose to inject the pitch histogram information here because of the comparable number of dimensions of the latent space and the histogram. Empirically, an additional fully-connected layer was needed to gradually project the dimensions of the output to 1, and for training to converge.

4.2.4 Training

Training the network requires positive (singer A better than or similar to singer B) and negative pairs (singer B better than A) of singer inputs. In our training, the ground truth label is 1 for positive pairs and 0 for negative pairs.

The loss function we minimize is the comparative loss which is a function of the probability output of the network and the binary ground-truth label, as given in equation 4. We use the Adam optimization algorithm [24]. The learning rate is 0.0001. The batch size is 10. Maximum number of epochs is set to 250, though early stopping based on training loss with patience of 5 epochs is employed for training termination. Back-propagation is car-

ried out through the twin-net arms. We choose the model that shows minimum loss in the validation set.

4.2.5 Prediction

The preference judgment value, i.e. the preference metric D from equation 3 of the twin network lies between -1 and 1. If this value is >=0, it implies singer A is preferred over singer B, thus the verdict is 1, and vice versa.

After all the pairwise comparisons, the singers can be rank-ordered according to the aggregate scores of each singer, given by the BWS score defined as

$$B = \frac{n_{best} - n_{worst}}{n} \tag{7}$$

where n_{best} and n_{worst} are the number of times the singer is marked as preferred and not preferred respectively, and n is the total number of times the singer appears.

4.3 Evaluation Metrics

We use three kinds of metrics to evaluate the performance of the framework with respect to the human ground-truth labels as described in section 4.1:

Pair prediction accuracy: This is defined as the percentage of input singer pairs for which the preference prediction from the network is correct.

Pearson's Score Correlation: This is the correlation between the machine BWS scores and human BWS scores.

Spearman's Rank Correlation: This is the correlation between the machine and human annotated singer rank-orders based on the respective BWS scores.

5. EXPERIMENTS AND RESULTS

The inter-judge correlation between ratings from music experts is 0.82 [4], which means that experts do not always agree with each other, and there is, in general, an upper limit of the achievable performance of any machine-based singing quality evaluation.

5.1 Twin-Net vs. Hybrid Twin-Net

We test our hypothesis that twin neural network can be applied for the task of learning singing quality preference in pairwise comparisons of singers to predict rank-ordering of singers. We train the twin-network and the hybrid-twin network on the 25,280 *same song* singer pairs from the training set of dataset 1. Since the two singers sing the same sequence of words, the twin arms focus on learning the discriminatory characteristics from the input representations which lie in the differences in the prosodic properties such as pitch harmonics of the two singing renditions. The hybrid network further helps in this process as the pitch histogram provides a direct singing quality discriminatory representation, as discussed in section 3.

From Table 2, we see that the both the twin-networks are able to converge on the training dataset with a high pair prediction accuracy and score correlation with humans. This validates our hypothesis and technique of the adaptation of a Siamese network for preference-based judgment and hence rank-ordering of singers. We also observe that the hybrid-twin network outperforms the twin-network on the test set from dataset 1. This implies that conditioning the network on pitch histogram frees degrees of freedom to model non-pitch related information via the network.

Dataset	%Accuracy		Pearson Corr.	
	Twin	Hybrid	Twin	Hybrid
Train	88.3	81.3	0.91	0.82
Validation	73.8	73.3	0.63	0.62
Test Dataset 1	72.7	76.1	0.61	0.68

Table 2. Performance of twin-network and hybrid twin-network in terms of pair classification accuracy and Pearson correlation between machine BWS scores and human BWS scores. All correlation values are statistically significant with $pvalue \ll 0.05$.

5.2 Comparison with Prior Studies

The prior studies that are closest to this work are the ones by Gupta et al. [11] and Pati et al. [25]. In the former, the authors studied various hand-crafted features to generate rank-ordering of singers, such as pitch histogram-based absolute measures and inter-singer distance based relative measures. They also performed late-fusion of these ranks to get a good correlation with human annotations. Pati et al. trained a supervised regression DNN model that uses mel spectrograms of pitched wind instruments as input representation to predict their subjective human scores.

In this experiment, we compare the performance of our proposed hybrid twin-network against the relative measures performance of [11]. Both these techniques involve same-song pair comparisons, and hence are conceptually similar. Additionally, we train the absolute score prediction network of [25] on our dataset. The ground-truth, in this case, are the raw human BWS scores of every singer that was provided with this dataset. This prediction network is similar to the absolute measures prediction from [11] in the sense that both involved direct assessment of singers. Finally, in late-fusion, we compute the average of the rank-order obtained from our hybrid twin network and the absolute score prediction network, similar to [11].

In Table 3, we see that the proposed hybrid-twin network performs better than the relative measures of [11]. Moreover, hybrid-twin outperforms the absolute score prediction network. This implies that pairwise comparisons in combination with pitch histogram representation results in better modeling of singing quality, than hand-crafted features. The late-fusion performances are comparable.

The inter-singer distances of relative measures in [11] compare the features from one singer with that of the rest of the singers in the dataset singing the same song. Thus, the major drawback of this method is that the relative measures will make sense only if all the singers are singing the same song. Moreover, for a new unseen song, there needs to be a large number singers singing that song for the thresholds designed for relative measures to be reliable. The above drawbacks make the relative measures highly song dependent. Moreover, any new test singer needs to be compared to all the singers in the dataset to get a reliable ranking. This becomes computationally cumbersome with increasing size of dataset. In the next sections, we show how our proposed framework overcomes these drawbacks.

Gupta et al. [11]		This work	
Framework	**Corr**	**Framework**	**Corr**
Relative Measures	0.64	Hybrid Twin-network	0.68
Absolute Measures	0.48	Absolute score prediction network [25]	0.62
Late-Fusion	0.71	Late-Fusion	0.71

Table 3. Comparison of the Spearman's rank correlation performance of the proposed hybrid twin network on dataset 1 with that from a recent previous work on the same dataset. All correlation values are statistically significant with $pvalue \ll 0.05$.

Framework	%Accuracy	Pearson's Score Corr	Spearman's Rank Corr
Twin-net	65.9	0.39	0.41
Hybrid twin-net	77.7	0.63	0.65

Table 4. The performance of twin-net and hybrid twin-net models on unseen songs from test dataset 2. The models are trained on the same song input training pairs from dataset 1. All correlation values are statistically significant with $pvalue \ll 0.05$.

5.3 Performance on Unseen Songs

To test the performance of the trained model on unseen songs, we evaluate its performance on test dataset 2 (Table 1). These songs and singers were not present in training set. The dataset consists of 90 same song singer pairs. From Table 4, we observe that the hybrid twin net outperforms the twin-net by a significant margin. This shows that the pitch histograms are a powerful representation of singing quality that reduces the dependency of the network on the identity of the song, thus confirming that our proposed framework can reliably evaluate unseen songs.

5.4 Comparing Different Songs

We further test if our proposed framework can compare singing vocals of different singing content. For this, we train the hybrid-twin net singer pairs singing same as well as different songs, for which we use the 102,080 ordered singer pairs of the training dataset (Table 1). In Table 5, we observe the performance of this model on the different song singer pairs from both test dataset 1, where the songs are seen by the trained model, and test dataset 2, where the songs are not seen by the trained model. Rank-ordering singing vocals with different-song singer-pair inputs (Table 5, row 1 and row 3) shows comparable results to same-song singer-pair comparisons (Table 3 row 1 and Table 4 row 2). Moreover, when rank-ordering is done using both different-song and same-song pair comparisons, the results on unseen songs (Table 5, row 4) significantly outperforms that from the same-song pair trained model (Table 4 row 2). This experiment shows that our proposed preference-based framework is able to learn discerning properties of singing quality such that given any two singers singing the same or different songs, it learns to choose the better singer.

5.5 Effect of Number of Comparisons

BWS method is known to be able to reliably rank-order with fewer number of comparisons. We tested this idea by reducing the number of paired comparisons in the test set, while ensuring that each singer appears at least once. Out of the 780 pairs, we randomly selected x number of unique pairs three times, and calculated the average of the performance of the three random trials. These average values of Spearman's rank correlation are plotted in Figure 4, where the number of pairs selected ranged from 550 to all of the 780 pairs. We observe that for a reduction of 30% in the number of pairs for comparison, there is a very small drop in the correlation value, approximately 6%. This computational advantage will become more significant when the size of the dataset increases.

Test Dataset	No. of diff. songs singer pairs	No. of same songs singer pairs	%Accuracy	Pearson's Score Corr	Spearman's Rank Corr
1	600	0	72.3	0.64	0.64
	600	180	72.7	0.65	0.65
2	100	0	77	0.68	0.68
	100	90	78.6	0.70	0.73

Table 5. Performance of hybrid twin network trained on the same and different song input pairs. All correlation values are statistically significant with $pvalue \ll 0.05$.

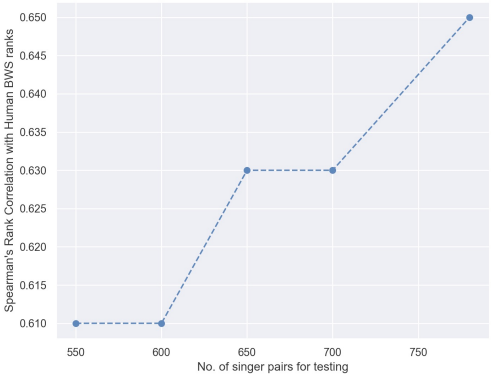


Figure 4. Spearman's rank correlation as the number of pairwise comparisons is reduced.

6. CONCLUSIONS

In this work, we propose a preference-based framework in which we adapt the twin neural network (Siamese) such that given two input singers, it learns to choose the better singer. We incorporate structural changes in the Siamese network framework such as preference metric instead of distance metric and comparative loss instead of contrastive loss, so that it is able to learn a preference instead of similarity. We show that with a few pairwise comparisons, this modified Siamese network effectively gives a reliable rank-order of singers. We also incorporate the musically relevant pitch histogram representation in a hybrid twin network framework, which shows to provide reliable singing quality predictions in a singer and song independent way on unseen data.

7. 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.

8. REFERENCES

[1] C. Gupta, "Comprehensive evaluation of singing quality," *PhD Thesis, National University of Singapore*, 2019.

[2] 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, 2012.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] J. M. Oates, B. Bain, P. Davis, J. Chapman, and D. Kenny, "Development of an auditory-perceptual rating instrument for the operatic singing voice," *Journal of Voice*, vol. 20, no. 1, pp. 71–81, 2006.

[9] 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.

[10] ——, "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.

[11] 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.

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

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

[14] A. Marley, T. N. Flynn, and V. Australia, "Best worst scaling: theory and practice," *International Encyclopedia of the Social & Behavioral Sciences*, vol. 2, no. 2, pp. 548–552, 2015.

[15] J. Louviere, I. Lings, T. Islam, S. Gudergan, and T. Flynn, "An introduction to the application of (case 1) best–worst scaling in marketing research," *International Journal of Research in Marketing*, vol. 30, no. 3, pp. 292–303, 2013.

[16] Y. Zhang, B. Pardo, and Z. Duan, "Siamese style convolutional neural networks for sound search by vocal imitation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 429–441, 2018.

[17] Y. Zhang and Z. Duan, "Visualization and interpretation of siamese style convolutional neural networks for sound search by vocal imitation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2406–2410.

[18] A. Gresse, M. Quillot, R. Dufour, V. Labatut, and J.-F. Bonastre, "Similarity metric based on siamese neural networks for voice casting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6585–6589.

[19] C.-i. Wang and G. Tzanetakis, "Singing style investigation by residual siamese convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 116–120.

[20] M. Panteli, R. Bittner, J. P. Bello, and S. Dixon, "Towards the characterization of singing styles in world music," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 636–640.

[21] K. Lee and J. Nam, "Learning a joint embedding space of monophonic and mixed music signals for singing voice," *arXiv preprint arXiv:1906.11139*, 2019.

[22] Y. Niu, D. Huang, Y. Shi, and X. Ke, "Siamese-network-based learning to rank for no-reference 2d and 3d image quality assessment," *IEEE Access*, vol. 7, pp. 101 583–101 595, 2019.

[23] 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.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] 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.