LEARN BY REFERENCING: TOWARDS DEEP METRIC LEARNING FOR SINGING ASSESSMENT

Huan Zhang[1] **Yiliang Jiang**[2] **Tao Jiang**[2] **Peng Hu**[2]
[1] School of Music, Carnegie Mellon University, Pittsburgh, US
[2] Tencent Music Entertainment, Shenzhen, China
huanz@andrew.cmu.edu

ABSTRACT

The excellence of human singing is an important aspect of subjective, aesthetic perception of music. In this paper, we propose a novel approach to tackle Automatic Singing Assessment (ASA) task through deep metric learning. With the goal of retrieving the commonalities of good singing without explicitly engineering them, we force a triplet model to map perceptually pleasant-sounding singing performance closer to the reference track compared to others, and thus learning a joint embedding space with performance characteristics. Incorporating mid-level representations like *spectrogram* and *chroma*, this approach takes advantage of the feature learning ability of neural networks, while using the reference track as an important anchor. On our designed testing set that spans across various styles and techniques, our model outperforms traditional rule-based ASA systems.

1. INTRODUCTION

Automatic Singing Assessment (ASA) deals with the task of assessing singing performances based on audio recordings. Ever since the development of Karaoke, a popular entertainment form and practice means for singers, there has been a high demand for ASA systems that's able to judge the excellence of singing performance just like human experts do.

However, ASA is not an easy task. Singing quality is often judged with respect to professional standards, where music experts rate singing performances based on their music knowledge and perceptual appeal. These dimensions include basic, objective criteria such as vowel quality (proper pronouciation of lyrics), accuracy of pitch and rhythm. Meanwhile, higher-level, subjective dimensions like singers formant, dynamics and expressiveness, techniques like vibrato and breathing are also taken into account.

Based on these criteria, some ASA systems compare a singing performance with a reference such as a professional singing performance [1,2] or melody contours [3,4], and thus place more emphasis on accuracy and intonation. On the other hand, unreferenced ASA systems aims to evaluate singing quality based on only the performance itself, addressing voice-related characteristics like voice formants or expressiveness. However, all these rule-based evaluation systems from hand-crafted indicators can easily be song- and style-dependent, resulting in poor generalizability. What's worse is that sometimes human perception standards seem to conflict with each other. For example, a good vibrato technique implies pitch instability, and an ASA system that's focused on intonation will fail on certain songs while human can easily perceive the trick. Given the vast dimensions of our perception space of singing quality, we seek data-driven, deep learning solutions.

In this work, we aim to tackle the ASA task through metric learning. With the goal of retrieving the commonalities of good singing without explicitly engineering them, we force the model to map perceptually pleasant sounding performances closer to the reference track compared to others, and thus learning a joint embedding space with performance characteristics. Section 2 reviews the related works. In Section 3, we first present our audio representations that incorporate multi-channel features. Then we introduce our proposed triplet network in comparison to other skeleton models for assessing singing quality. Experiment results and discussions are shown in Section 4, where our proposed architecture achieved the highest correlation with human perception of singing rating on the mixed testing set, when comparing with existing singing assessment algorithms.

2. RELATED WORK

Currently, the majority of literature on ASA systems, whether referenced or unreferenced, largely focused on extraction of perceptually motivated features such as f_0 pitch sequences [3], f_0 pitch histogram [5], vibrato rate (periodic fluctuation of f_0) [6]. The main issue of these hand-crafted features is that they only reflect specific aspects of the singing. Thus, some of the systems [1, 7, 8] will then feed features into simple machine learning regressors or classifiers which predict ratings that take advantage of multiple features.

With the rise of deep learning, deep neural network is

Attribution: H. Zhang, Y. Jiang, T. Jian and P. Hu, “Learn by Referencing: Towards Deep Metric Learning for Singing Assessment”, in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.,* Online, 2021.

found to outperform traditional methods in terms of feature learning. To our knowledge, the only work that tackles ASA task though end-to-end deep learning without feature engineering step is [9], which takes in a mid-level time-frequency representation of singing clips and evaluate singing as binary (good-poor) classification with a bi-dense neural network. However, this approach does not incorporate the song reference melody, and thus only assesses the discernible features which are independent of the particular singer or melody.

How to make use of the song reference information while taking advantage of deep learning? We observe that the technique of deep metric learning has recently attracted much attention in various MIR tasks. As a method of learning discriminative features by measuring similarities from samples, most existing studies used Siamese and Triplet networks to correlate among samples while using shared weights [10]. In [11–13], these architectures has been applied to tasks of Singer Identification, Cover Song Detection and Singing Style Investigation. What's more, [14] assesses woodwind instruments performances via a joint embedding network, confirming the feasibility of learning a performance-reference joint latent space. In [15], an attempt with deep metric learning was used for assigning scores for singing tracks, but the foundation was laid with a classification network.

Compared to previous works, the novelty of our work is in two folds: **1)** We take a deep metric learning approach for singing assessment, which utilizes reference tracks (original, accompaniment) while not being constrained by it. **2)** Besides classical time-frequency representation of audio, we propose a set of mid-level audio representations which concerns pitch, tempo, timbre and so on. At the end, we construct a style-mixture testing set that comprehensively evaluates systems' ability in assessing different dimensions of singing, and provide a detailed discussion.

3. METHODOLOGY

3.1 Audio Representation

Given that we are not explicitly engineering perceptually motivated features, it is important to present neural network with a comprehensive view of the audio data. In such comprehensive task like singing assessment, it is unusual to judge only the pitch accuracy of a performance and ignore tone production, or only care about hitting the right beat but not pronunciation of the lyrics. Thus, for input features, we employ five channels of 2-dimensional audio representations that concern with musical dimensions such as pitch, rhythm, timbre:

i *Log-Mel Spectrogram* **(Spec)**: We extract the Log-Mel Spectrogram for each *3s* segment with a hop size of 512. Given 16kHz sampled audio, the resulting representation contains 94 frames and 96 bins.

ii *Chroma* **(Chroma)**: Chromagram gives us the pitch class profiles for the clips. Note that in order to achieve the same dimension (96) with other channels, we give 6 bins (96 *bins* / (12 *pitch classes*)) for each pitch class.

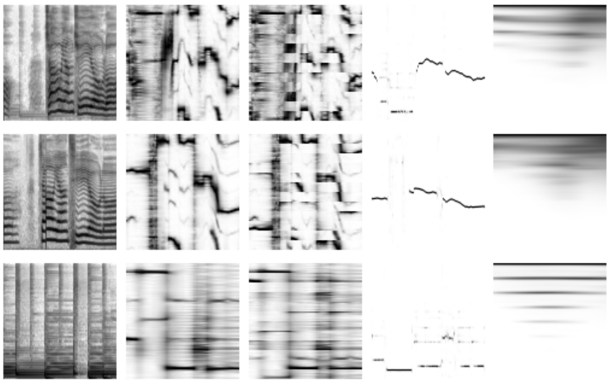

Figure 1. **Top**: *Spec, Chroma, TChroma, F0, Tempo* 2D visualization of a clip that's labeled as 'good'; **Mid**: another clip that's labeled as 'bad' at the same timestamp; **Bottom**: features from their corresponding accompaniment clip.

iii *Tonal Shifted Chroma* **(TChroma)**: From a functional harmony perspective, pitch *G* and *C* are closer while *C*♯ and *C* differs more. Thus, in this channel we take inspiration from [16] and rearrange the rows of chromagram by circle of fifths, in the hope that the features are more sensitive to non-harmonic mistakes.

iv *F0* **(F0)**: As seen in Section 2, f_0 is the most crucial feature for assessing intonation. Pitch extraction algorithm Crepe [17] is used to obtain an activation matrix of estimated pitch.

v *Tempogram*(**Tempo**): We also extracts the cyclic tempogram [18] of the clip, representing the estimated tempo that evolves over time.

We adopted Librosa's [19] implementation for all of the audio features above except F0. In [20], it was studied that a short voiced sequence (3-5 sec.) is sufficient for assessing singing quality. Thus, for all of input audio we extract the features above from $3s$ clips, with 96×94 in dimension concatenated together like a 5-channel image. See Figure 1 for comparison of the 5 channel representations of a pair of $3s$ clips and their corresponding reference clip.

3.2 Architectures

For the schematic model for embedding, we adopted the CNNSA (CNN + Self Attention) model proposed in [21, 22]. Originally designed for music-tagging task, the architecture achieved compelling results in learning local characteristics and temporal representations via interpretable attention maps.

Structurally, the model employs a front-end / back-end division of deep neural network for MIR task that was first proposed by [23] (Figure 2): The front-end consists of a 7-layer CNN with (3 × 3) filters with residual connections [24], aiming to extract local information such as timbre and pitch. The back-end utilizes stacks of self-attention layers to achieve temporal summarization like rhythmic patterns, melodic contours, and chord progressions. Here, the self-attention layers are BERT [25] encoders where Q, K, V are feature maps obtained from front-end.

3.2.1 *Baseline: Direct Score*

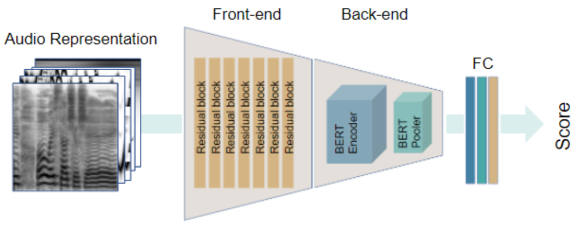

Figure 2. Baseline architecture of deriving score from audio representation, without referencing or contrasting.

As shown in Figure 2, our baseline model is a direct regression model that learns a mapping between audio representations and the labeled score directly. No reference is used. For the architecture, the five-channel audio representation is fed into the schematic *CNN* + *self attention* model mentioned in 3.2, with 3 fully connected layers at the end to output a score that measures the excellence of singing.

3.2.2 *Delta Spectrogram*

The "delta spectrogram" (`Delta`) model is a natural extension of the baseline architecture with a reference clip: Given the audio representation, we subtract it from the audio representation of reference clip. Given most of our audio inputs have time-frequency attributes, a singing clip misaligned with pitch or rhythm will be reflected in its delta with reference clip. In comparison with other architectures, this method is not attempting to learn a joint latent space, and can be viewed as extracting local similarities at the front of the pipeline.

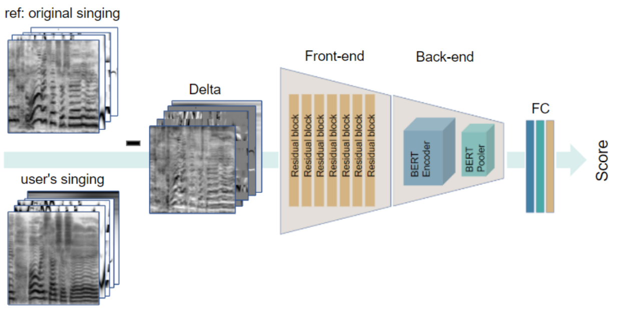

Figure 3. The `Delta` model, where the notion of distance with anchor is incorporated as delta at the front of the pipeline.

3.2.3 *Triplet:*

The triplet model is our main proposed architecture for learning a joint embedding space with singing characteristics. The objective of this architecture is to learn a mapping from audio representation to an embedding space where good singings are closer to reference while poor singings are further away in this space. We implement this idea via a triplet network shown in Figure 4, which takes in (reference (as anchor), good singng (as positive), poor singing (as negative)).

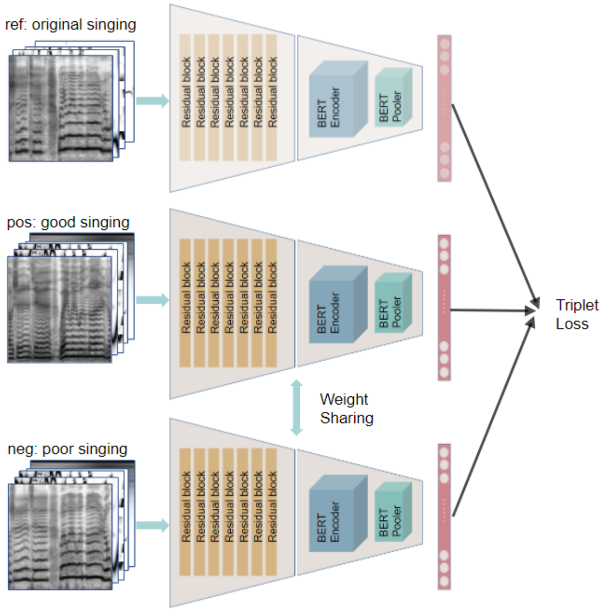

Figure 4. Metric learning architecture.

All three towers have the same architecture, but the two towers of positive and negative singing clips enabled weight-sharing while the reference tower doesn't. This is because the performance tracks and reference tracks came from different audio domains, and thus benefits from projecting to different embedding spaces. Afterwards, a 128-dimensional embedding for the performance is learnt through the model, where we optimize toward Triplet Margin Loss with Cosine Similarity as distance.

The Triplet Margin Loss is computed by

$$L(a, p, n) = max\{D(a, n) - D(a, p) + \alpha, 0\}$$

where a, p, n denotes $anchor, positive, negative$ respectively, and α is the margin. For the distance metric D we utilize the inverse of Cosine Similarity, as a larger similarity represents smaller distance:

$$D(x, y) = -\frac{x \cdot y}{\|x\| \cdot \|y\|}$$

Given the 128-dimension output of the model, the final assessment output is the cosine similarity between the embedding of the given performance and the embedding of its reference track. Thus, the score assesses the proximity of the performance towards reference. For the triplet architecture, we utilized both original singing and accompaniment as reference tracks, and the experiments are denoted by `Rori` (reference with original) and `Racc` (reference with accompaniment).

3.2.4 *Embedding Direct Score*

With the belief that the joint latent space encodes the characteristics of commonalities of good and poor singing, we also trained a embedding direct score (`EmbDirect`) that takes the assessment as a downstream task after a singing embedding is learnt from Section 3.2.3.

Specifically, for a given clip, we take the `Rori` model pre-trained in previous section and output the clip embedding of 128 dimension. In this system, we train three fully connected layers with ReLU activation to regress and match the labelled score. In inference, the final score is obtained from the audio clip through the embedding and final output layers directly. Thus, this is also an unreferenced system.

3.3 Data

The training data are collected from the solo singing clips without accompaniment from *** [1]. The singers are volunteers who were the common users of this application.

For reference tracks used in metric learning, we perform experiments with both *accompaniment* tracks and *original singing* tracks, which are the pure singing voice from the published version of song. The accompaniments are purchased by *** from music production studios, and original singing tracks are source-separated by Spleeter [26].

After making sure that the reference track and solo singing track are exactly aligned, WebRTC vad [27] is used to detect the voiced segments from the singing clip. Each *3s* segment from the voiced segment, along with the referenced track clip on the same position, is extracted into audio representation in Section 3.1. Our data preprocessing pipeline is shown in Figure 5.

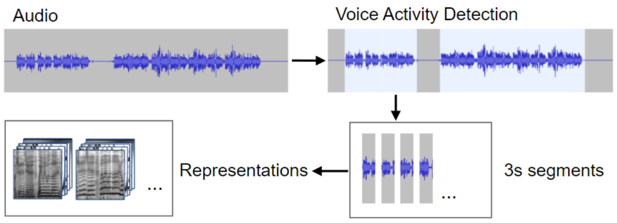

Figure 5. Preprocessing pipeline

Another question is on how we obtain the "good" and "poor" singing for deep metric learning. For the clips used in our training, a quality score within $[0, 100]$ was labeled by the company's contractor employees. Note that these quality scores are very rough as they come from multiple people's standards without a detailed listening, and there is no guarantee for their coherency. We take all clips with scores $\geq$ 80 as "good" and $<$ 40 as "poor". In the mashup, we randomly align a "good" performance and a "bad" performance of the same song to form a contrasting pair.

We believe the quality score is not exact in modeling the excellence of singing, but gives a rough direction on what's good and what's poor. Meanwhile, the weak labelling is actually advantageous to our exploration as they represent a general perception and prevents our models to overfit to any specific assessment standard.

In total, we obtained 15487 *3s* singing clips pairs, and an equal number of positive and negative data are exactly aligned. These clips are obtained from 1240 full length recordings and from 102 songs. All clips were resampled to 16K Hz. In terms of genre, the songs roughly consists of 75% of Chinese pop with a variety of tempo and style (published after 2000), 15% of folk songs, 5% of rock, and 5% of other genres. There are no jazz or classical singing styles in the training set.

The testing set we designed will be introduced in Section 4.

3.4 Experiments

We trained our models on 3 NVIDIA V100 GPUs on a single machine. For the CNN models, the optimizer of the triplet loss is ADAM [28] with a learning rate 10^{-6}. For `EmbDirect` system, we used learning rate 10^{-7} as it only trains 3 linear layers. We used batch size of 128 and trained our models for a maximum of 200 epochs with early stopping based on validation loss with patience of 5 epochs. We choose $\alpha = 1$ for the margin in Triplet loss.

Song	Genre	Character	BPM	Ori.singing
茧 (Cocoon)	pop	lyrical, spans over large range	78	nasal, genderless voice; young male
夜夜夜漫长 (Night is Long)	electronic, pop	rhythmical, low register, rap-like, energetic	98	husky, hoarse; middle-aged male
白月光与朱砂痣 (White Moonlight and Cinnabar Nevus)	pop	simple melody and slow harmonic rhythm, small range	89	deep, melodious; middle-aged female
Love Story	country, pop	high tempo, high note density, short notes and syllables	120	silvery, sweet; young female
今夜草原有雨 (It's Gonna Rain in the Steppe Tonight)	Chinese folk-style, pop	high register, a lot of vibrato, extremely expressive	62	operatic, soprano; middle-aged female

Table 1. Characteristics and style analysis of the songs in testing set.

4. EVALUATION AND DISCUSSION

Our testing set consists of 5 songs with different styles, that spans over genres like pop, electronic, country and folk, as summarized in Table 1 [2]. The dataset consists of a mix of songs that spans over various registers, tempo, techniques and even language; they also attracts different cultural and age groups. For each of the 5 songs we subjectively choose 9 different performances with various quality, creating a testing set of 45 recordings.

[1] https://***app.com/ *** is a Karaoke application.

[2] bpm of the songs are estimated using Madmom [29]

	Configuration	Reference Used	Song1	Song2	Song3	Song4	Song5	Mix
Baseline	Direct	None	0.785	0.223	0.472	0.528	0.253	0.417
Proposed	Delta	Original Track	0.718	0.383	0.684	0.288	0.276	0.430
Proposed	Rori	Original Track	**0.912**	**0.860**	0.521	0.839	0.480	**0.652**
Proposed	Racc	Accompaniment Track	0.635	0.708	0.853	0.663	-0.222	0.459
Proposed	EmbDirect	None	0.861	0.503	0.256	0.487	**0.753**	0.533
Histogram	peakBW	None	0.875	0.581	0.714	0.872	0.281	0.626
Histogram	peakConc50	None	0.651	0.836	0.822	0.736	0.407	0.520
Histogram	binning	None	0.874	0.776	0.688	**0.882**	-0.206	0.521
DTWDist	pitch	MIDI	0.812	0.732	**0.907**	0.324	0.068	0.589
DTWDist	volume	Energy Sequence	0.723	0.524	0.761	0.468	0.432	0.467
DTWDist	rhythm	MIDI	0.361	0.542	0.606	0.042	0.043	0.279

Table 2. Pearson's correlation between human scoring and algorithm scoring. Correlations among each individual song and mixture of all songs are shown.

4.1 Subjective Ground Truths

For each of the 45 recordings, we asked 5 professionals (graduates from music conservatoire with 10+ years of performance training) to assign them scores based on the quality of singing. Overall, the scores reached an average inter-judge correlation of 0.78, and thus we consider them as valid ground truth. For evaluation of the systems, we computed the Pearson's correlation coefficient between the output scores from the algorithms and the human judge's mean score.

4.2 Objective

As mentioned in Section 2, there are plenty of rule-based methods for singing assessment tasks. We implemented our version of two existing methods: the pitch histogram measures proposed in [5] and feature (pitch, dynamics, rhythm) evaluation method proposed in [3, 30].

4.2.1 Histogram

The pitch histogram method is an unreferenced system that computes a series of attributes for evaluating unaccompanied singing. With bins of 100 cents within an octave, pitch histogram represents the distribution of pitch values in a performance. Good singings will usually have sharp peaks on specific note values, while poor singings will have dispersed distribution as they sing out of tune. Thus, a series of measurements relating to the spread of peaks in the histogram can be used for evaluation. We computed scores like kurtosis, skewness, peakBW, peakConc50, peakConc110, kMeans,binning as described in [5], and listed the best 3 measurements in Table 2.

4.2.2 DTWDist

As specified in [3, 30], we also implemented traditional Pitch-based Rating, Volume-based Rating, Rhythm-based Rating. In general, these methods computes the **DTWDist** (Dynamic Time Wraping Distance) between users' performances with reference (MIDI note sequence for pitch and rhythm, recording energy sequence for volume), and more similar sequences indicates better singing. We also presents the performance of these ratings in Table 2.

4.3 Results Discussion

Table 2 shows the comparative performances for all systems. We are able to make the following observations:

i **Performances of deep learning systems.** Among all deep learning systems, our Rori system proved to correlate the most with human perception of singing ranking. The models (Direct, Delta) that don't incorporate reference tracks were outperformed, showing that learning a joint latent space of singing quality indeed helps with our assessment goal. Between the deep metric learning systems, we noticed that Racc doesn't perform as good as Rori, since the accompaniment track as anchor does not provide details on singing, but only helps with judging the rhythm and tonality.

ii **Learnt embeddings are more robust to song variations.** For each individual song, there is at least one system that reaches 0.75 of correlation with human perception, but on the mixture of 5 songs the maximum correlation we can obtain is 0.65. This confirms the concern mentioned in Section 1 , that it's difficult for ASA systems to evaluate different songs on the same scale. The pitch histogram measures, while performing great among the rating within each songs, suffers when we cross compare performances from different songs: The 'spikes' within the histogram are influenced by the number of pitches used, and some songs are naturally going to achieve a higher score in their metrics. In comparison, deep metric learning approaches are more robust. See also Section 4.4 for more demonstrations.

iii **Ability to evaluate on more nuanced techniques.** Song 5 is a Chinese folk style piece that demands singing techniques such as vibrato. It's difficult to hand-craft perception motivated features for such techniques, and neither of the traditional feature-based method perform well on this piece. The

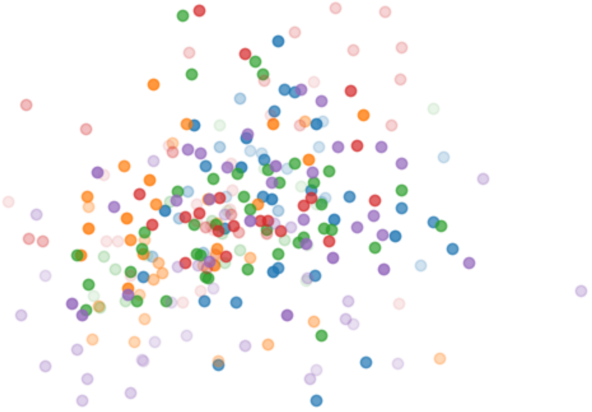

Figure 6. PCA Projection of clip embeddings from the testing set. 5 colors correspond to songs, and transparency is scaled with the human evaluation score of performance, where higher scores are darker. Best viewed in color.

deep learning based models `Rori`, `EmbDirect`, however, both outperform traditional methods on this song while retaining high performance on other songs. This confirms the ability of deep neural network to model performance-related features.

iv **Does good implies similar?** The best performing model for Song 5 is actually `EmbDirect` system (Section 3.2.4), that assesses the singing through features learnt from the good-poor metric space without computing the similarity with original singing directly. Thus, we speculate that there is this conceptual gap between "Good" and "Similar": For a given piece, there are multiple ways of performing it nicely, and they don't necessarily need to be similar to the original. The `EmbDirect` system demonstrates that the learnt joint embeddding space encodes singing characteristics and can be used to tell good or poor singing apart - to learn by reference while not being constrained by it.

4.4 Embedding Space Visualization

In Figure 6, we take all the 3s clips from 45 singing performances in the testing set, and project their 128-dimensional embeddings by PCA. The embedding is trained from our `Rori` configuration that obtained the best cross-song result in Table 2, contrasting good-poor performance with original singing. 5 songs are distinguished by colors, while transparency represents the assessment from ground truth human judgements, where poorer singing has a higher transparency.

The embedding space visualization supports our observations in Section 4.3: In the vocal embeddings, song differences are largely eliminated, while the clips from higher scored performances tends to cluster in the middle and lower scored performances are dispersed. This demonstrates that, regardless of songs and references, the embeddings capture certain universal characteristics in distinguishing excellency of singing voice.

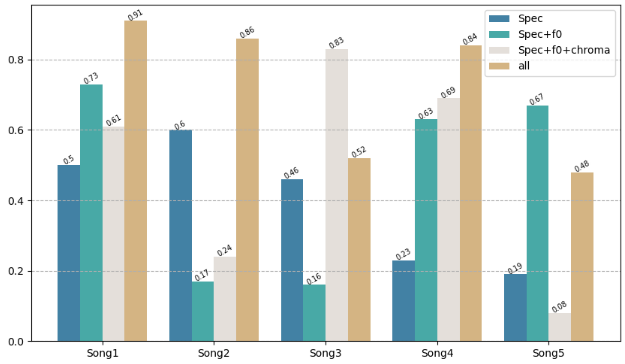

Figure 7. Comparison between different input audio representations.

4.5 Ablation Study for Input Audio Representation

To demonstrate the effect of our multi-channel audio representation described in Section 3.1, ablation experiments were performed to show that the combination of audio representation indeed achieved better assessment results. For the `Rori` architecture, we performed experiments using 4 combinations: $Spec$, $Spect+f_0$, $Spec+f_0+chroma$, and all 5 representations together.

Figure 7 shows the performance of different audio representation input on 5 songs respectively. Overall, the input that utilized all 5 channels achieved the best result. Given that Convolutional Neural Network is still one of the most popular architecture for 2D audio feature learning nowadays, this idea of presenting a multi-channel view to deep networks may be applied to other interesting tasks.

5. CONCLUSION AND FUTURE WORK

This paper presents a novel approach of automatic singing assessment task via metric learning. Through training a triplet model that anchors at a reference track of the performance, we were able to learn a joint embedding space where characteristics of good and poor singing were extracted. Comparative experiments were performed on a designed testing set that evaluates assessment systems across variety of singing styles and techniques. Results demonstrate that the proposed system outperforms baseline and feature-based assessment systems in cross-song ratings when correlates with human judgments.

Given the intrinsic subjective aspect of human perception of music performance, singing assessment as well as broader music assessment has been little investigated through deep learning approach. Our work demonstrates that it's possible to direct deep neural network in learning performance related characteristics via comparing weakly labeled data. Future explorations may expand on this "learn by reference" idea with other paradigm such as contrastive learning [31], or apply to neighboring domains like instrumental assessment. Also, we wish our experiments with multi-channel audio representations would facilitate more explorations in musically-motivated input design.

6. REFERENCES

[1] C. Gupta, H. Li, and Y. Wang, "Perceptual evaluation of singing quality," in *Asia-Pacific Signal and Information Processing Association*, 12 2017.

[2] P. Lal, "A comparison of singing evaluation algorithms," in *INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP*, vol. 5, 01 2006.

[3] W. Tsai and H. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Trans. Speech Audio Process.*, vol. 20, no. 4, pp. 1233–1243, 2012. [Online]. Available: https://doi.org/10.1109/TASL.2011.2174224

[4] 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*, 2013, pp. 744–748.

[5] C. Gupta, H. Li, and Y. Wang, "Automatic evaluation of singing quality without a reference," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 11 2018, pp. 990–997.

[6] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP*, vol. 4, 01 2006.

[7] Bar, Bozkurt, O. Baysal, and D. Yüret, "A dataset and baseline system for singing voice assessment," in *13th Int. Symposium on Computer Music Multidisciplinary Research*, 2017.

[8] J. Böhm, F. Eyben, M. Schmitt, H. Kosch, and B. Schuller, "Seeking the superstar: Automatic assessment of perceived singing quality," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 1560–1569.

[9] 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.

[10] M. Kaya and H. . Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, p. 1066, 2019.

[11] K. Lee and J. Nam, "Learning a joint embedding space of monophonic and mixed music signals for singing voice," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Delft, Netherland*, 2019.

[12] M. Stamenovic, "Towards cover song detection with siamese convolutional neural networks," 2020. [Online]. Available: https://arxiv.org/abs/2005.10294

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

[14] J. Huang, Y.-N. Hung, A. Pati, S. K. Gururani, and A. Lerch, "Score-informed networks for music performance assessment," in *21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[15] T. Tan, "Singing evaluation based on deep metric learning," in *Proceedings of the 2019 3rd International Symposium on Computer Science and Intelligent Control*, ser. ISCSIC 2019. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3386164.3389096

[16] R. Abecidan, M. Giraud, and G. Micchi, "Towards Custom Dilated Convolutions on Pitch Spaces," International Society for Music Information Retrieval Conference (ISMIR 2020), Late-Breaking Demo Session, 2020, poster. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02959676

[17] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[18] P. Grosche, M. Müller, and F. Kurth, "Cyclic tempograma mid-level tempo representation for musicsignals," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 04 2010, pp. 5522 – 5525.

[19] B. McFee, V. Lostanlen, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, J. Mason, D. Ellis, E. Battenberg, S. Seyfarth, R. Yamamoto, K. Choi, viktorandreevichmorozov, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Hereñú, F.-R. Stöter, P. Friesch, A. Weiss, M. Vollrath, and T. Kim, "librosa/librosa: 0.8.0," Jul. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3955228

[20] T. Nakano, M. Goto, and Y. Hiraga, "Subjective evaluation of common singing skills using the rank ordering method," in *9th International Conference on Music Perception and Cognition*, 2006.

[21] M. Won, S. Chun, and X. Serra, "Toward interpretable music tagging with self-attention," *arXiv preprint arXiv:1906.04972*, 2019.

[22] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of cnn-based automatic music tagging models," in *Proc. of 17th Sound and Music Computing*, 2020.

[23] J. Pons, O. Nieto, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," *Proceedings ofthe International Society for Music Information Retrieval Conference (ISMIR)*, vol. abs/1711.02520, 2018.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 06 2016, pp. 770–778.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.

[26] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020, deezer Research. [Online]. Available: https://doi.org/10.21105/joss.02154

[27] A. B. Johnston and D. C. Burnett, *WebRTC: APIs and RTCWEB Protocols of the HTML5 Real-Time Web*. St. Louis, MO, USA: Digital Codex LLC, 2012.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[29] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "madmom: a new Python Audio and Music Signal Processing Library," in *Proceedings of the 24th ACM International Conference on Multimedia*, Amsterdam, The Netherlands, 10 2016, pp. 1174–1178.

[30] W. Tsai, C. Ma, and Y. Hsu, "Automatic singing performance evaluation using accompanied vocals as reference bases," *Journal of Information Science and Engineering*, vol. 31, no. 3, pp. 821–838, 2015. [Online]. Available: http://www.iis.sinica.edu.tw/page/jise/2015/201505_04.html

[31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607. [Online]. Available: http://proceedings.mlr.press/v119/chen20j.html