AUTOMATIC SINGING EVALUATION WITHOUT REFERENCE MELODY USING BI-DENSE NEURAL NETWORK

Ning Zhang, Tao Jiang, Feng Deng, Yan Li

Kuaishou Technology Co. Beijing

ABSTRACT

Automatic singing evaluation without reference melody has long been a difficult problem. This paper aims to pilot a novel data driven approach to tackle this artistic problem. We constructed a large scale dataset and designed an innovative Bi-Dense neural network which can address this task efficiently. Though the singing evaluation is quite a subjective task and depends a lot on listeners' preferences, we showed that a specific group has consistency on the singing evaluations, and it is possible to train a model to learn the subjective preferences of this group. In this paper, a large amount of singing clips and corresponding human gradings were collected. And an elaborate designed Bi-DenseNet was trained to discriminate the good singings from the poor ones. The experiments demonstrated the proposed network performs better than the existing networks for singing evaluation task.

Index Terms— singing evaluation, Bi-DenseNet

1. INTRODUCTION

Singing evaluation has long been thought as a complex task, which inherently is subjective and listener dependent [1]. Generally, a listener's preference on the singings is influenced by one's experience, character or something virtual we called taste. In the past, the singing evaluation has been mostly achieved by human music experts, who are responsible to tell the public if a song is well sung by a specific person, which typically is a professional singer. It is also common that these experts disagree with each other in their evaluations. Nowadays, besides professional singers, everyone can sing a song and publish it on the social networks. Hundreds of millions of singing works have been published everyday on various website and apps. On one hand, it's impossible for human experts to listen all these songs and score them. On the other hand, the preferences of the human experts can not represent those of ordinary people. Therefore, an automatic singing evaluation system can be very useful for karaoke scoring, singing recommendation, and entertainment singing contest. More importantly, this work is also a great trial to bridge the field gap between vocal artists and deep learning techniques.

Singing evaluation has not received too much research endeavor for a long time. A few investigators from perceptual domain and MIR domain made significant contributions to the study of solo singing performance from the auditory-perceptual aspect and acoustic aspect. Wapnick and Ekholm quantified the correlations between12 perceptual items (including intrinsic qualities, execution abilities, diction and others) and the overall evaluation scores [2]. Oates Jennifer et.al [3] tested the similar items and developed an auditory-perceptual rating instrument for operatic singing voice. Garnier Mava et.al [4] defined the notion of voice quality in Western lyrical singing and investigated significant and objective criteria to characterize it, from both cognitive and acoustic points of view. Nakano et.al [5] explored the criteria that human subjects used in judging singing skill and the stability of their judgments. They focused on ordinary, common person's singing, and their mutual evaluations. They showed that subjects depend more on objective common features (such as tonal and rhythmical stability) rather than subjective preference. They also found that only a short sequence (3-5 sec.) is sufficient for judging good/poor. Cao Chuan *et al.* [6] [1] studied the subjective criteria for untrained singers's singing voice quality evaluation, including intonation accuracy, rhythm consistency, timbre brightness and vocal clarity.

In addition to these perceptual investigation, various automatic singing evaluation methods have been proposed in the past years. The existing automatic singing evaluation methods can be classified according to if the reference melody is used. The reference melody based methods extract various acoustic features, including pitch, volume, rhythm, timbre and others, from the singing clips. Then compared these features to those of the reference basis, which is often the vocal track from original music recording (CD or VCD) [7] [8]. However, in many cases, we have no access to the reference vocal tracks. Human subjects can consistently evaluate the singing for unknown melodies [5]. This suggests that their evaluation utilizes easily discernible features which are independent of the particular singer or melody. Nakano et.al [9] proposed an automatic singing skill evaluation method for unknown melodies. They showed that pitch interval accuracy and vibrato are useful acoustic features for singing skills evaluation of unknown singers and melodies. These hand-crafted features can only reflect specific aspects of the singing skills, which cannot represent the overall singing levels.

This paper aims to build an automatic singing evaluation

model without reference basis. The study object is the ordinary people's judgements on common peoples's singing clips. With in this scope, we constructed a large singing dataset, named SIE dataset, and the human grades for each singing clips in this dataset. We proposed a novel Bi-DenseNet which represents singing voice features efficiently. The experiments showed that the proposed Bi-DenseNet outperforms the existing neural networks on the singing evaluation task.

Section 2 introduces the collected dataset and singing evaluation background. Section 3 describes the proposed Bi-Dense neural network. Section 4 is the experiments and analysis. Section 5 concludes this paper.

2. THE SIE DATASET

2.1. Data collection

We collected the singing clips tagged 'solo singing' from Kwai[1]. Totally $30,570$ clips published between Jun. 1st. and Aug. 31, 2018 were obtained. These clips were sung and published by the common users of Kwai application. Their recording devices varied from user to user. Most users recorded with their cell phones. Others may use different microphones. Besides, the recording environments were also varied, ranging from at home, outdoor, in car to on bus. We went through these audio clips, and removed those accompanied by background music or other instruments. $19,478$ pure solo singing clips are remained. The time length of these clips ranges from 10 seconds to 3 minutes. All these clips were resampled to 16K Hz, and saved as mono channel 16 bit wave files.

2.2. Human grading

Each of the above pure solo singing wave files were presented to 10 listeners independently. All these listeners are our workmates and of college degrees or above, with ages 20 to 30. These listeners were required to listen to each singing clip completely and classify the presented clip into *good*, *intermediate* and *poor*. In order to let the listeners be familiar with the distribution of the whole dataset and understand their tasks better, firstly 2700 singing clips were sampled from the whole dataset randomly and listeners listened to these clips and tried to classify them into three classes. Then the total 19,478 singing clips were presented to listeners one by one, and the listeners gave their classification results for each data. In addition, the genders of the singers for each clips were labelled by other labelers.

Each of the 19,478 singing clips has 10 classification labels from the 10 listeners. To measure the consistency of the ten labels for each clips. We computed the variational ratio for each clip. The variational ratio is a simple measure of statistical dispersion of a nominal distributions, and is defined as the proportion of cases which are not in the mode category [10]. The variational ratio for each clip was computed and then averaged over the whole dataset. The averaged variational ratio is 0.3277, which implies that for a given singing clip, over $2/3$ listeners are inclined to give the same classification labels. We conjectured that maybe it is possible to use the 'average label' to represent the group preference on each clip. Following this supposition, we assigned a score to each label class. The scores for *good*, *intermediate* and *poor* are 5, 3, 1, respectively. We can compute the coefficient of variation (CV) [10] for the ten labels of each clip. The CV, which is defined as the ratio of standard deviation to the mean, is a standardized measure of dispersion of a frequency distribution. The average CV over the whole dataset is 0.3320, which indicates that the scores' deviation from the mean is relatively low (comparing to when the scores are evenly dispersed, CV in this case is greater than 0.5). Hence we can use the average score to denote the score of a specific singing clip. To further eliminate confusions, the clips with average score equal to or greater than 4.0 are labelled as *good*, and the clips with average score equal to or less than 2.0 are labelled as *poor*. Other clips were treated as *others*.

2.3. Dataset construction

We dropped the data labelled as *others*, and only data labelled as *good* and *poor* were kept. As proved in [5]. A 3-5 seconds short sequence is sufficient for judging good/poor. Considering there are maybe blanks at the start/end or middle part of a song, we extended this time length to 10 seconds. To standardize this dataset, we segmented all the singing clips using 10-seconds sliding window, with 2.5s hop size. More than 33 thousands segments were obtained. The data distribution of this set is shown in Table 1. This dataset, named as the SIE (SingIng Evaluation) dataset, is imbalanced across the classes and genders.

Based on the SIE set, we constructed a balanced dataset (SIE-22k), shown in Table 2. The SIE-22k dataset consists of a training set and a testing set. All the segments in training set and testing set come from different singing clips. These data are evenly distributed in classes and genders. Note that only $female$ and $male$ singing clips were remained for sake of clarity.

3. MODEL DESIGN FOR SINGING EVALUATION

3.1. Preliminary

More and more researches use deep learning models for audio analysis tasks [11].Though recurrent neural networks (RNNs) are powerful in modeling sequential data. Convolutional neural networks (CNNs) are usually used to learn efficient representations from audio waveforms [12] or fft-spectrograms [13] [14]. Despite that the CNNs were initially created for the image classification tasks, we have seen some successful cases that

[1]https://www.kwai.com. Kwai is a mobile social application on which everyone can publish short video works

gender	good	poor
female	5657	6257
male	5878	11302
female&male	44	63
unidentified	12	4425

Table 1. The number of segments in the whole singing evaluation (SIE) dataset. The gender of singers are also shown. 'female & male' means there are both female and male singers in this clip. 'unidentified' means that it's hard to identify the gender of the singers according to the singing clips. Most 'unidentified' singers are children.

set	gender	good	poor
training	female	4396	4396
	male	4396	4396
testing	female	1235	1235
	male	1235	1235

Table 2. The number of segments in the balanced singing evaluation dataset (SIE-22k).

the existing CNN architectures can be directly transferred to audio processing tasks. For example, Hershey *et al.* used Alex Net and VGGs for audio events detection and audio tagging [15]. Naoya and Yuki used DenseNet for audio source separation [16].

In this paper, we proposed an innovative neural network architecture named Bi-DenseNet to learn to discriminate the good singings from the poor singings. The Bi-DenseNet was created on the basis of DenseNet, which was proposed by Gao Huang *et.al.* in [17]. The DenseNet takes the insights of the skip connection to the extreme, in which the output of a layer is connected to all the subsequent layers in the module. The DenseNet can also be viewed as multi-scale feature extractor that features in lower scale are used to generate higher scale features and finally features of all scales are combined together to compute the classification results.

However, the original DenseNet was well-designed for image processing task. Its convolutional layers are efficient to extract image features. Directly applying this architecture to music data does not make sense. In this paper, we tailored the DenseNet to the singing evaluation task. As stated in [6], the most relevant features for singing evaluation include pitch, rhythm, timbre. The proposed Bi-DenseNet was designed to account for these multi-scale temporal and spectral features of singings.

3.2. Bi-DenseNet

The model architecture of Bi-DenseNet is shown in Fig. 1. The 10 seconds 16K Hz singing segments were transformed to time-frequency domain through the STFT with hann window of 1024 point, and hop of 512 point, resulting to a 513×313 magnitude spectrograms, which were taken as input to the network. The Bi-DenseNet is composed of the input convolution block, K Bi-Dense blocks and transition layers, and the output layers.

Input Conv. blocks. The timbre of singers are mostly featured by the pitch and its harmonic partials, which mostly spread along the frequency axis in the time-frequency domain. The rhythm and structure of a singing related to local patterns in short and long time-scales. [14] and [18] designed the convolutional filters using domain knowledge to extract the timbre and temporal features. Here, we designed an efficient convolution layer to extract features from the input magnitude spectrograms. This layer is constituted with a set of horizontal and vertical rectangular convolutional filters to accomodate the temporal and spectral features of singing segments. The vertical filters are of shape $M \times 3$, with $M \in [128, 196, 256]$, 12 filters for each shape. The horizontal filters are of shape $3 \times N$, with $N \in [64, 128, 256, 320]$, 16 filters for each shape. Strides of these filters are devised to be 2×1. Stride *2* in frequency axis can reduce the frequence dimension of input spectrograms by two times without losing any information. But stride> 1 in time axis does not make sense, it will distort the sequence and loss rhythm information. Hence the the stride 1 is used for time axis. The outputs of the different convolutional filters are combined together by concatenating the channel axis and then fed to the Bi-Dense blocks for further process.

Bi-Dense blocks. This block consists of L composite functions, each of which comprise three consecutive operations: batch normalization, rectified linear unit(ReLU), and convolutions. To better resolve the feature output above, we placed parallel convolution filter banks in the convolution layer of each composite function. Specifically, each convolution layer comprises 12 horizontal filters of shape 1×5 and 12 vertical filters of shape 5×1. The outputs of these convolutional banks in each layer are concatenated together and input to all the subsequent composite functions. Strides for all the convolutions are 1×1. 20% dropout is applied after each convolution layers.

Transition layers. As devised in the original DenseNet, the transition layers consist four consecutive operations: batch normalization, ReLU, a 3×3 convolution, and an average pooling of 2×2. The number of feature-maps is reduced by a factor of r. We chose $r = 0.5$ in this paper.

Output layers. The output from the last Bi-Dense block is global averaged and then passed to a linear transformation layer followed by 2-way softmax activation.

4. EXPERIMENTS AND ANALYSIS

To prove that the singing evaluation task can be addressed via the proposed Bi-Dense neural network, we devised a set

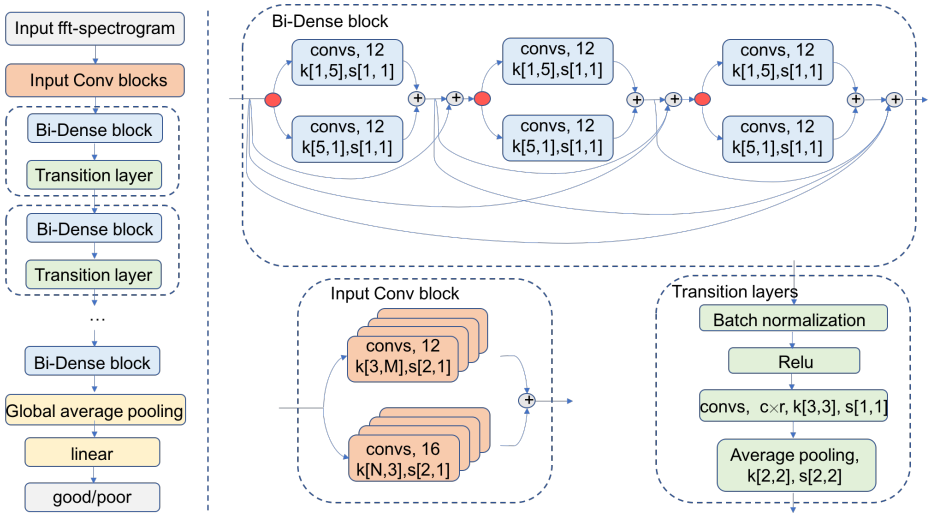

Fig. 1. Model architecture. Left: Overview of the blocks in Bi-DenseNet model. Top-right: The Bi-Dense block. Bottom-Middle: The input convolutional blocks. Bottom-right: The transition layer. The red circles in the Bi-Dense block correspond the operations 'Batch Normalization-Relu', k and s denote the kernel and stride size respectively. c in the transition layers denotes the number of input channels, r is the reduction factor.

Models & Param.		Precision (%)		Recall (%)		Acc (%).
		good	poor	good	poor	
VGG-E	-	50.00	0.0	100.0	0.0	50.00
DenseNet-10	K=1, L=8	96.73	72.78	63.40	97.85	80.63
DenseNet-27	K=4, L=6	96.36	78.14	72.79	97.25	85.02
Bi-DenseNet-10	K=1, L=8	81.82	85.10	85.83	80.93	83.38
Bi-DenseNet-27	K=4, L=6	92.74	86.81	85.83	93.28	**89.55**

Table 3. The classification results on the SIE-22k test set. *Acc.* denotes the classification accuracy. K is the number of Dense blocks in the model, L is the number of composite functions in each block.

of comparison experiments on the SIE-22k dataset. The fft-spectrograms of the 10 seconds segments were computed as above. Then the fft-spectrograms were compressed by applying the element-wise dynamic range compression function $f(x) = log(1 + C \cdot x)$, where $C = 10,000$ is a constant controlling the amount of compression [19]. Global mean and standard deviation were computed over the training set. The compressed fft-spectrograms were normalized to zero mean and unit variance. The loss function used in this paper was cross entropy. The Bi-DenseNet was trained using momentum gradient decent method with Nesterov momentum 0.9 [20]. The initial learning rate is set to 0.1, and is divided by 10 at the 200k, 300k, and 500k training steps. Batch size for training was set to 4. The model was trained on the training set of SIE-22k, and tested on the test set. Different parameters of the model were explored.

For comparision, we also tested the performance of existing neural networks, the Vgg-E net and the original DenseNet. These models are trained under the same condition, except that the initial learning rate for VGG-E was set to $1e-3$. The classification results for these models are shown in Table 3. Precisions and recalls for the good and poor classes as well as the classification accuracy are shown. From this table, we can see that the VGG-E model failed on the singing evaluation task, all the tested clips were classified as *good*. The original DenseNets achieved relatively better results. However, they performed unequally on *good* and *poor* classes. Significant numbers of *good* samples were classified as *poor*, resulting in low recalls on the *good* class. The proposed Bi-DenseNets perform more evenly among two classes, and outperform the original DenseNets under the same parameters in terms of the overall classification accuracy. The Bi-DenseNet-27 achieved the highest overall classification accuracy among all the tested networks.

5. CONCLUSION

This paper aims to build an automatic singing evaluation model which can discriminate the good singings from the poor singings. We firstly collected large amounts of singing clips sung by common people and the corresponding grades evaluated by ordinary listeners. A balanced dataset SIE-22k were constructed for the training and testing. We proposed a novel Bi-DenseNet which is suitable for audio processing and can extract efficient features from singing spectrograms. The experiments demonstrated that the proposed Bi-DenseNet performs better for singing evaluation task than the existing networks, including the VGG-E and the original DenseNet.

6. REFERENCES

[1] Chuan Cao, Ming Li, Jian Liu, and Yonghong Yan, "An objective singing evaluation approach by relating acoustic measurements to perceptual ratings," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[2] Joel Wapnick and Elizabeth Ekholm, "Expert consensus in solo voice performance evaluation," *Journal of Voice*, vol. 11, no. 4, pp. 429–436, 1997.

[3] Jennifer M Oates, Belinda Bain, Pamela Davis, Janice Chapman, and Dianna Kenny, "Development of an auditory-perceptual rating instrument for the operatic singing voice," *Journal of Voice*, vol. 20, no. 1, pp. 71–81, 2006.

[4] Maëva Garnier, Nathalie Henrich, Michèle Castellengo, David Sotiropoulos, and Danièle Dubois, "Characterisation of voice quality in western lyrical singing: From teachers' judgements to acoustic descriptions," *Journal of interdisciplinary music studies*, vol. 1, no. 2, pp. 62–91, 2007.

[5] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga, "Subjective evaluation of common singing skills using the rank ordering method," in *Ninth International Conference on Music Perception and Cognition*. Citeseer, 2006.

[6] Chuan Cao, Ming Li, Jian Liu, and Yonghong 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.

[7] Wei-Ho Tsai and Hsin-Chieh 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.

[8] Wei-Ho Tsai, Cin-Hao Ma, and Yi-Po Hsu, "Automatic singing performance evaluation using accompanied vocals as reference bases.," *J. Inf. Sci. Eng.*, vol. 31, no. 3, pp. 821–838, 2015.

[9] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga, "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.

[10] Raymond H Myers, "Elementary applied statistics," 1974.

[11] Emilia Gómez, Merlijn Blaauw, Jordi Bonada, Pritish Chandna, and Helena Cuesta, "Deep learning for singing processing: Achievements, challenges and impact on singers and listeners," *arXiv preprint arXiv:1807.03046*, 2018.

[12] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M Schmidt, Andreas F Ehmann, and Xavier Serra, "End-to-end learning for music audio tagging at scale," *arXiv preprint arXiv:1711.02520*, 2017.

[13] Jiyoung Park, Jongpil Lee, Jangyeon Park, Jung-Woo Ha, and Juhan Nam, "Representation learning of music using artist labels," *arXiv preprint arXiv:1710.06648*, 2017.

[14] Jordi Pons, Olga Slizovskaia, Rong Gong, Emilia Gómez, and Xavier Serra, "Timbre analysis of music audio signals with convolutional neural networks," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 2744–2748.

[15] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "Cnn architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–135.

[16] Naoya Takahashi and Yuki Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on*. IEEE, 2017, pp. 21–25.

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks.," in *CVPR*, 2017, vol. 1, p. 3.

[18] Jordi Pons and Xavier Serra, "Designing efficient architectures for modeling temporal features with convolutional neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2472–2476.

[19] Sander Dieleman and Benjamin Schrauwen, "End-to-end learning for music audio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6964–6968.

[20] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, 2013, pp. 1139–1147.