Singing Voice Detection Using Multi-Feature Deep Fusion with CNN

Xulong Zhang[1], Shengchen Li[2], Zijin Li[3], Shizhe Chen[4], Yongwei Gao[1], and Wei Li[1,5]

[1] School of Computer Science, Fudan University, Shanghai, 201203, China
[2] Institute of information photonics and optical communications, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[3] Department of Musicology, China Conservatory of Music, Beijing, 100101, China
[4] Department of Music Engineering, Shanghai Conservatory of Music, Shanghai, 200031, China
[5] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, 201203, China

Abstract. The problem of singing voice detection is to segment a song into vocal and non-vocal parts. Commonly used methods usually train a model on a set of frame-based features and then predict the unknown frames by the model. However, the multi-dimensional features are usually concatenated together for each frame, with little consideration of spatial information. Hence, a deep fusion method of the Multi-feature dimensions with Convolution Neural Networks (CNN) is proposed. A one dimension convolution is made on feature dimensions for each frames, then the high-level features obtained can be used for a direct binary classification. The performance of the proposed method is on par with the state-of-art methods on public dataset.

Keywords: Convolution Neural Network (CNN), Multi-feature Fusion, Deep Learning, Singing Voice Detection (SVD)

1 Introduction

In the field of Music Information Retrieval (MIR), singing voice detection (SVD) is to locate the vocal portions in a piece of music, which can be seen as a useful preprocessing step for a variety of MIR tasks, such as singer identification [1], singing voice separation [2], singing voice melody transcription [3], query by humming [4], lyrics transcription etc. The main difficulty of SVD mainly comes from the extent of vocal tone diversity.

The issues of SVD are usually addressed through traditional statistical methods [5], such as Gaussian Mixture Models (GMM), neural networks and support vector machines (SVM) [6], Hidden Markov Model (HMM) [7] and etc. Eyben et al. [8] proposed the data-driven approach based on Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) for Voice Activity Detection (VAD) in speech. The main advantage of the LSTM model is the ability to

H. Li et al. (eds.), *Proceedings of the 7th Conference on Sound and Music Technology (CSMT)*, Lecture Notes in Electrical Engineering 635,
https://doi.org/10.1007/978-981-15-2756-2_4

model long range dependencies between input series. For the successful use of LSTM on numerous research areas, Lehner et al. [9] introduced the LSTM to SVD. There are up to 111 audio features used in feature representation. The use of LSTM achieves the state-of-art performance on the two publicly available datasets (Jamendo [6] and RWC [7]). Leglaive [10] added bi-directional structure on LSTM that takes the past and future temporal context into account on the presence/absence of singing voice. In Schlüter's work [11], Convolutional Neural Networks (CNN) model on Mel spectrograms is used to design the singing voice detection system. The CNN model has been demonstrated powerful to learn invariances taught by data augmentation in other fields.

Except the design of classifier in singing voice detection, another important module is feature representation. These studies [12][13][14]are based on features mostly come from speech processing field, such as Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Cepstral Coefficients (LPCC). However, these features may not be good enough to distinguish the singing voice from background music. Rocamora and Herrera [15] found that MFCC and their derivatives are the most appropriate features, the accuracy was around 78.5%.

Regnier and Peeters [12] presented a method to detect vocal segments within an audio track based on two specific characteristics of the singing voice, vibrato and tremolo. In [16], Lehner et al. optimized the MFCC features with manually tuned parameters, they achieved 82.36% accuracy after a temporal smoothing. In another work [17], three new features are designed by Lehner. Though the results are competent to other works, the process of feature extraction is too complicated.

Single feature cannot fully describe the audio features, and concatenation features together simply may lead incompatible between different features or highly dependent on a certain dimension of the feature sets. Therefore, we can treat each feature dimension as a one dimensional space. Due to the space relation information, CNN is used to learn invariant features, which can relieve the complexity of manual design features and make different dimensions' feature more compatible. In addition to CNN models for the fusion features, a preprocess of singing voice separation is applied to get more focus on the vocal part and a post-process of temporal smoothing to modify the obvious temporal exception.

The rest of this paper is structured as following. An overview of our SVD system is presented in section 2. Experiments and results on common benchmark datasets are then presented and discussed in section 3. Finally, some conclusions are proposed in section 4.

2 Proposed SVD system

The architecture of the proposed system uses singing voice separation (SVS) as the preprocess step to get vocal signal, then follows a traditional bag-of-frames approach: a machine learning technique (CNN) is applied on a set of features computed on successive frames of the incoming vocal signal. The output of the classifier is then further being temporal smoothed to localize musical segments

that contain singing voice. The overview of our SVD system is shown in figure 2, and different building blocks are described in detail below.

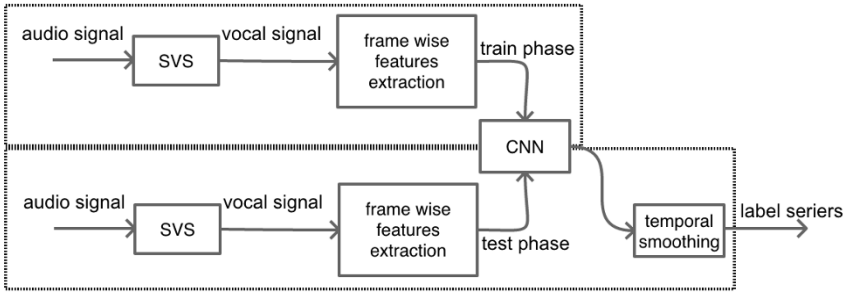

Fig. 1. The proposed SVD system overview

2.1 Singing Voice Separation

Singing voice separation (SVS) is taken as a preprocess step, which can split the mixed music signal into vocal portions and accompaniment portions. SVS method is based on the "REPET-SIM" method [18] with modifications and extensions: FFT windows are overlapped by 1/4, instead of 1/2. Non-local filtering is converted into a soft mask by Wiener filtering. This is similar in spirit to the soft-masking method used by [19], but is a bit more numerically stable in practice. Nevertheless, the split of accompaniment portions and vocal portions is hard to be completely separated due to the strong accompaniment, the separated vocal contains a little harsh noise. After SVS we save three versions of audio signal on the same dataset, then we can compare their results with the same classifier.

2.2 Feature Extraction

This section briefly describes the chosen features. There are many features proposed for the singing voice detection problem. To Among these features, MFCC [20], LPCC [21] and Chroma [22] are examined in this paper. The three features were chosen for feature fusion cause that they describe mixed audio at three different aspects. MFCC has been widely used in many speech and audio recognition tasks [15] and MFCC can represent the timbre of the audio signal. LPCC features are calculated by introducing the cepstrum coefficients (CCs) to the Linear Predicative Coding (LPC) parameters. LPCC feature reveals the nature of the produced sound which is governed by the shape of the vocal tract. The Chroma features are a well-established tool for analyzing and comparing music data [23].

The audio signal is segmented into 40ms frames with an overlapping of 20ms. FFT is calculated on each frame with Hamming window. Most features were selected for their ability to discriminate vocal with music [24].

The features were calculated on short-scale frames stated above. 26 MFCC coefficients (without any energy coefficient), 12 LPCC coefficients and 12 Chroma coefficients were extracted from each frame. Finally, the combined feature vector has 50 dimensions (MFCC-26, Chroma-12, LPCC-12).

2.3 CNN for feature deep fusion and Classification

In this paper, we present a novel data-driven method for singing voice detection based on deep fusion of features with the CNN model. The motivation behind the use of CNN is the capacity to learn compositional representations in spatial, where invariants from the original feature spatial can be learned. It not only can extract deep features, but also can use these learned feature for the binary classification directly.

The proposed networks for SVD have an input layer which matches the size of the combined acoustic feature vectors, two one-dimension Convolutional layers, two one-dimension Max pooling layers, and dense layers to flatten the input to the output layer with a single sigmoid unit.

The input of the CNN is successive frames. The multi-features were extracted and then concatenate together to build the raw feature representation of each frame. Then there are two dimensional inputs: x axis represents the temporal series of frames while the y axis represents different dimensions of features. The one-dimension Convolutional will process on x axis where the flatten layer will combine the final extracted features of different frames. Connect them to the dense layer for output.

The networks are trained as classifier to output a voicing score for every frame in the value space of 0 and 1:1 indicating singing frame, 0 indicating no singing frame. The final step in predicting the audio frame at timestamp t is to take a softmax and then do a temporal smoothing over the outputs of the sequential model. The neural network topologies are shown in figure 2.3.

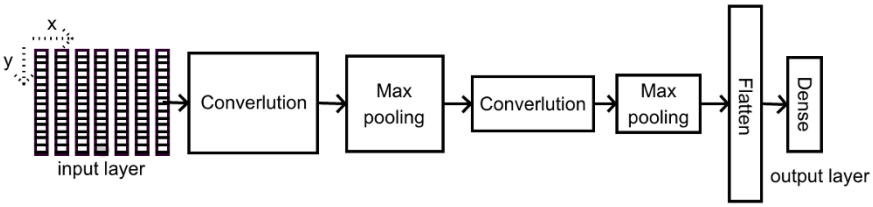

Fig. 2. The neural network topologies of the proposed CNN

2.4 Post Process

Due to the classifications vary significantly. The likelihood ratios vary wildly from frame to frame. This is in stark contrast to the out labeling data where the class labels stay the same for many successive frames. Given that the singing voice has continuity in a certain period, it is more reliable to accumulate the segment likelihood over a longer period of decision making.

In this paper, we proposed three methods for segmentation smoothing. The first one is the median filter to smooth the raw classification variable along the time dimension. The second method is to use the posterior probabilities obtained by a Hidden Markov Model of two states (vocal and non-vocal). The observation distributions are modeled by a mixture of 45 Gaussians, fitted with the Expectation Maximization algorithm. The best path of states is then deduced from the classifier output sequence with the Viterbi algorithm. The third one is the Conditional Random Field (CRF) that was used to learn the relation of the prediction and the ground truth on validation dataset.

3 Experiments and results

For comparison, we choose two public mainly used datasets for SVD. We also make common evaluation on accuracy, precision, recall and F1 measure.

We compare the performance of the difference features to determine how helpful they are and how to parameterize them, and then combine the best features. Besides, the best performance system on the two public datasets against our implemented LSTM under the same conditions (feature representation, preprocess and post-process) are also compared.

3.1 Benchmark Datasets

To our knowledge, there are two publicly available corpora with vocal activity annotations. One is Jamendo corpus and the other is RWC pop music dataset.

Jamendo Corpus [6] has a set of 93 songs, which constitute a total of about 6 hours of music. Each file was annotated manually into singing and non-singing sections by the same person to provide the ground thrust-data. As Jamendo Corpus had been split the 93 songs into 3 parts to generate train, test and valid sets. We use the same split dataset for training and testing as the compared relate work.

The RWC Popular Music Dataset [7] contains 100 pop songs, with singing voice annotations by Mauch. As this dataset commonly was used by 5-fold cross validation and use the average performance for compilation. The same split was done as the compared relate work.

3.2 Evaluation

In order to give a comprehensive view of the results, we compare model predictions with the ground truth labels to obtain true positives(TP), false positives

(FP) true negatives (TN) and false negatives (FN) over all songs in the test set. Besides, we also calculate the frame-wise accuracy, precision, recall and with F1 measure to summarize results. These metrics can be represented in below equations:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - measure = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

3.3 Results

In the first experiment, the mixed audio of music signal was split into vocal and accompaniment and the 13 MFCC coefficients were extracted as the audio feature to train the classifier without the post process on the prediction. We compare binary classification results on split vocal, the split accompaniment and the mix audio of music signal. The performance on the preprocessed data and mix audio is showed in figure 3.3. We use 'jamendo vocal' to denote the SVD result of vocal part after singing voice separation on Jamendo dataset. And with the 'jamendo mix' represent the result of Jamendo raw data without any preprocess of SVS. 'jamendo music' is the result of the music part after SVS of Jamendo dataset. The result labeled on RWC dataset was similar at Jamendo.

From the comparison of the classification results use different audio signal on two datasets. The use separated vocal is higher than the raw mix signal by 2% in accuracy. The accompaniment music signal is lower than the raw mix signal, so we can conclude that apply the preprocessing of singing voice separation can improve the final vocal detection performance. In the music, the accompaniment often strong and not only overlapped with the vocal in temporal it also intertwined with vocal in frequency. So do a preprocess of singing voice separation can degrade the influence by the accompaniment.

We compare the performance of different features that we choose in section 2.2. In the first experiment, three different features and their combinations (LPCC, Chroma, MFCC) are compared to classify the vocal and non-vocal segments using the deep CNN model separately. Through this experiment we want to check if the CNN can obtain more effectively information from the single features. The performance of different features on separating vocal part of RWC and jamendo dataset is shown in figure 3.3 and 3.3.

From figure 3.3 and 3.3, the Chroma feature is not fit for the vocal detection task. Although MFCC was very popular in the relate work such as in [16], in this experiment, MFCC performed normally. The LPCC has almost the best performance compared with the other two features. When the three features'

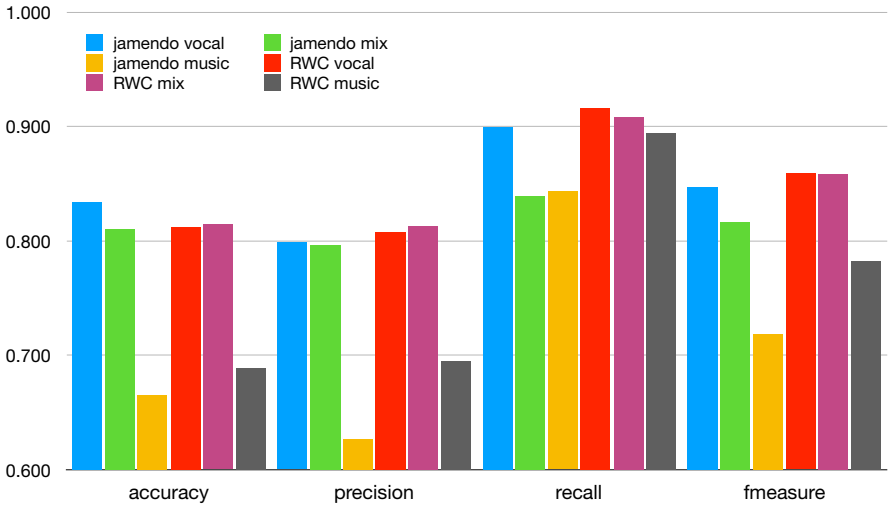

Fig. 3. Different audio data with singing voice separation as preprocess

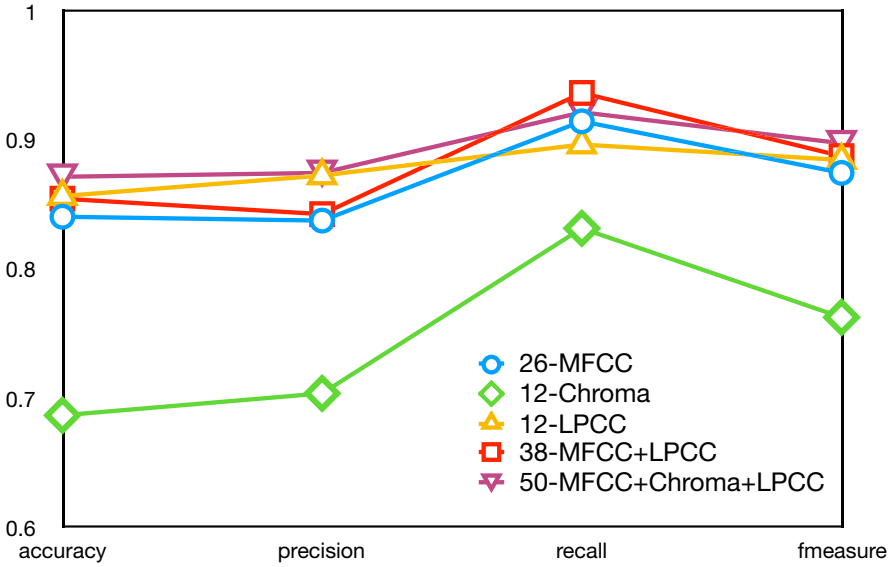

Fig. 4. Different features performance on RWC dataset

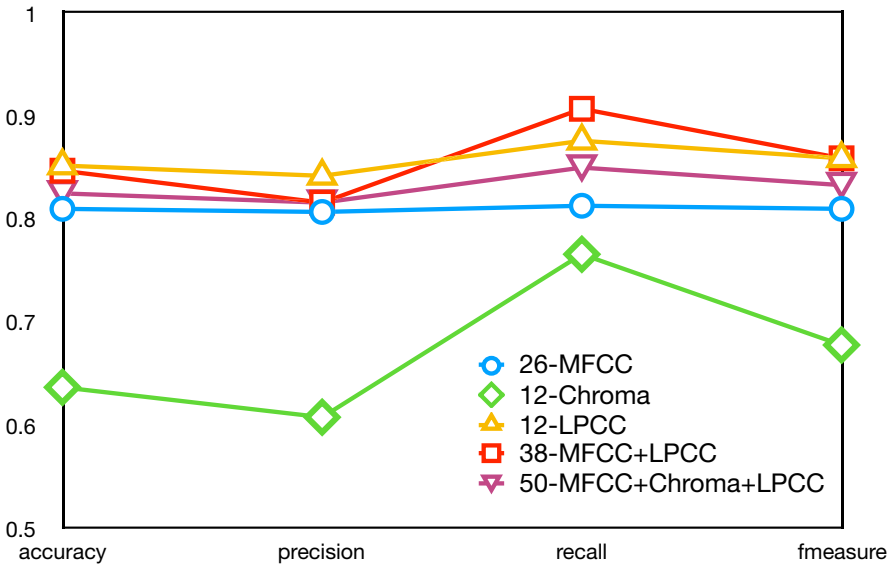

Fig. 5. Different features performance on Jamendo dataset

combination vector was chosen and feed to CNN model. CNN fusion different dimensions of the feature well. Therefore, the worst performance of Chroma feature does not affect the overall performance much.

In order to verify the effect of post-processing, the different temporal smoothing method were compared in the post process described in section 2.4. Firstly, median filtering was used for temporal smoothing and the number is decided by experiment on validating dataset. In the second, the HMM was used on the predict probability that HMM based method not need the fixed length frames window. The total series probability was used to train the HMM and then the prediction of the final segments boundary can be got. Due to HMM was used as an unsupervised model, it just uses the information of the classifier's prediction probability. Given a probability sequence, a 2-state HMM was trained on it, and at the end of the training process, run the Viterbi algorithm on the sequence to get the most likely state associated with each input vector. The segment boundary was found by HMM, and then the frames from each segment vote for the final label. The third is CRF with nearly the same as HMM, but the main difference is that CRF is a supervised model. The validation dataset was used to get the prediction and the ground truth, then train the CRF model. The trained CRF model is used for predicting the smoothing label series.

From the comparison result show in figure 3.3, the post process is necessary for singing voice detection. Compare with the blue line one without using post processing which labeled as 'Without postprocess', the performance can be improved by 4% by HMM. The performance of smoothing process of median fil-

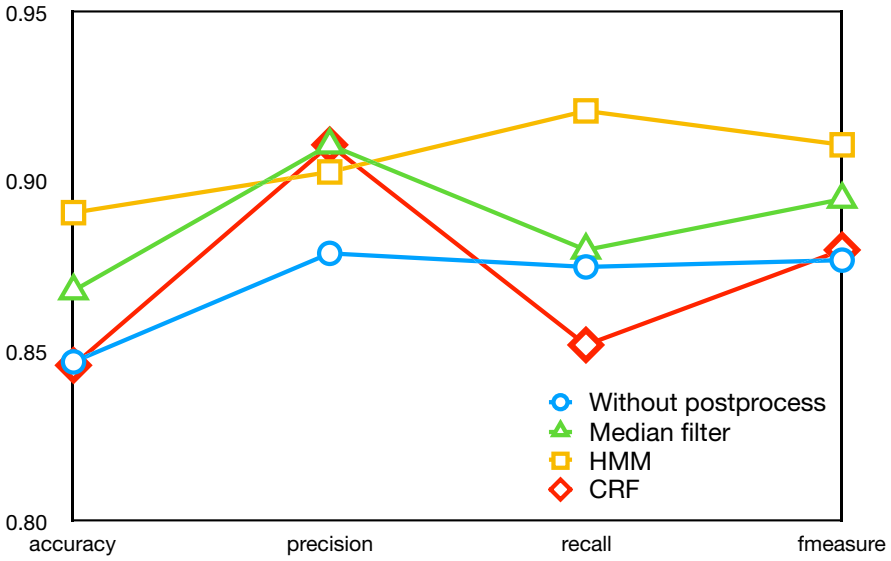

Fig. 6. With different temporal smoothing on RWC dataset

tering and CRF are both weaker than the case of HMM. Median filtering smooths the sequence in a fixed window and it leads the original boundary disappears and generates new fixed length segments. So use median filtering will produce more false positives, so the recall is getting smaller. As to the CRF model, it needs data sequence split into parts. Although there is no need to set each part of a fixed length in the training phase. When use the trained model need set the length, so there also has a boundary problem. If the length either too small or too long will lead to inconsistencies in training and testing.

On the public datasets we used, there are several works achieve the state of art on the task of singing voice detection. Finally, the proposed systems are compared with Ramona [6], Schlüter [11], Lehner-1 [17], Lehner-2 [16], Lehner-3 [9], Leglaive [10] on Jamendo corpus. And compare with Mauch [7], Schlüter [11], Lehner-1[17], Lehner-2[16], Lehner-3[9] for RWC pop dataset. The comparison results are presented in table 3.3 and table 3.3.

Our model is called proposed CNN in Table 1. A LSTM network is also used as a baseline system for comparison with the same preprocess and post process, which are called imLSTM.

Table 3.3. shows the comparison results on Jamendo dataset. We implement imLSTM and CNN by Keras and run it on GPU to get our results, while the other 6 results are reported in the related report on the public dataset Jamendo. The results demonstrate that on this dataset, Leglaive (uses the BLSTM-RNN) still keep the state-of-art best performance. For imLSTM with the combined three features, the F1 measure value is 0.796 which is lower than the Ramona's

Table 1. Proposed SVD System Compared with Others on Jamendo Corpus

	Accuracy	Precision	Recall	F1
Ramona	0.822	-	-	0.831
Schlüter	0.923	-	0.903	-
Lehner-1	0.882	0.880	0.862	0.871
Lehner-2	0.848	-	-	0.846
Lehner-3	0.894	0.898	0.906	0.902
Leglaive	0.915	0.895	0.926	0.910
imLSTM	0.795	0.897	0.716	0.796
Proposed CNN	0.859	0.917	0.796	0.853

SVM. But with proposed CNN model, the F1 measure gets an improvement with 5 percent. Although it has not yet reached the best performance, CNN is valid to fusion different feature dimensions to a compatible way and the result is better than the LSTM under the same conditions.

Table 2. Proposed SVD System Compared with Others on RWC Pop Dataset

	Accuracy	Precision	Recall	F1
Schlüter	0.927	-	0.935	-
Mauch	0.872	0.887	0.921	0.904
Lehner-1	0.875	0.875	0.926	0.900
Lehner-2	0.868	0.879	0.906	0.892
Lehner-3	0.923	0.938	0.934	0.936
imLSTM	0.868	0.902	0.887	0.894
Proposed CNN	0.890	0.911	0.912	0.911

Table 3.3. shows the comparison results on RWC pop dataset. On this dataset, the-state-of-art best results are kept by Lehner-3 (uses LSTM and well-design feature sets). For imLSTM with the combined three features, the F1 measure value is 0.894, it is on par with other 5 methods except the best results. Finally, the proposed CNN gets F1 measure value of 0.911, only latter than the-state-of-art best result. Compared to these two datasets, we can find that the performance on RWC is better than Jamendo. It may be because of data labeling difference between these two dataset. Jamendo corpus was labeled by one person, while the RWC pop dataset was labeled by a team. There may be some errors for manual labeling.

4 Conclusions

In this paper, a novel SVD system based on CNN with the fusion of multi-feature dimensions was proposed. In the SVD system, vocal is separated out of the mix audio signal, and CNN was used to fusion the different features dimension of the same frame. With a post processing of temporal smoothing, the performance of the proposed SVD based on CNN can be on par with the state-of-art performance on public dataset.

For future works, we will investigate the performance of CNN in more detail, such as analyzing the context learning behavior using time-frequency domain features or modulation spectrum features. Besides, we will learn compositional representations in spatial and temporal domain. Make combination of LSTM and CNN to the SVD. Furthermore, semi-supervised and active learning could be used to efficiently adapt the generic models presented in this paper to other tasks such as singer identification.

5 Acknowledgement

This work is supported by NSFC 61671156.

References

1. Kim Y, Whitman. B. Singer identification in popular music recordings using voice-coding features[C]. ISMIR. Paris, France, 2002.
2. Vembu S, Baumann. S. Separation of vocals from polyphonic audio recordings[C].ISMIR. London, UK, 2005.
3. Salamon e a J. Melody extraction from polyphonic music signals: Approaches, applications, and challenges[J]. IEEE Signal Processing Magazine. 2014.
4. Hsu e a C.-L. A tandem algorithm for singing pitch extraction and voice separationfrom music accompaniment[J]. IEEE Transactions on Audio, Speech, and LanguageProcessing. 2012.
5. Sonnleitner e a R. A simple and effective spectral feature for speech detection in mixedaudio signals[C]. DAFx'12. York, UK, 2012.
6. Ramona G R M., David B. Vocal detection in music with support vector machines[C].ICASSP. Las Vegas, NV, USA, 2008.
7. Mauch e a M. Timbre and melody features for the recognition of vocal activity andinstrumental solos in polyphonic music[C]. ISMIR. Miami, Florida, USA, 2011.
8. Eyben e a F. Real-life voice activity detection with lstm recurrent neural networks andan application to hollywood movies[C]. ICASSP. Vancouver, BC, Canada, 2013.
9. Lehner G W B., Bock S. A low-latency, real-time-capable singing voice detection-method with lstm recurrent neural networks[C]. EUSIPCO. Nice, France, 2015.
10. Leglaive R H S., Badeau R. Singing voice detection with deep recurrent neural networks[C]. ICASSP. South Brisbane, Queensland, Australia, 2015.
11. Schlüter J, Grill T. Exploring data augmentation for improved singing voice detectionwith neural networks[C]. ISMIR. Malaga, Spain, 2015.
12. Regnier L, Peeters G. Singing voice detection in music tracks using direct voice vibrato detection[C]. ICASSP. Taipei, Taiwan, 2009.

13. Pikrakis e a A. Unsupervised singing voice detection using dictionary learning[C].EUSIPCO. Budapest, Hungary, 2016.
14. Li X F W., Xue M. Reducing manual labeling in singing voice detection: An activelearning approach[C]. ICME. Seattle, WA, USA, 2016.
15. Rocamora M, Herrera P. Comparing audio descriptors for singing voice detection inmusic audio files[C]. Brazilian Symposium on Computer Music. San Pablo, Brazil,2007.
16. Lehner R S B., Widmer G. Towards light-weight, real-time-capable singing voice de-tection[C]. ISMIR. Curitiba, PR, Brazil, 2013.
17. Lehner G W B., Sonnleitner R. On the reduction of false positives in singing voicedetection[C]. ICASSP. Florence, Italy, 2014.
18. Rafii Z, Pardo. B. Music/voice separation using the similarity matrix[C]. ISMIR.Porto, Portugal, 2012.
19. FitzGerald D. Vocal separation using nearest neighbours and median filtering[C]. ISSC.Maynooth, Ireland, 2012.
20. You Y C W S.D., Peng S H. Comparative study of singing voice detection methods[J].Multimedia tools and applications. 2016.
21. Gupta H, Gupta D. Lpc and lpcc method of feature extraction in speech recognition system[C]. 6th International Conference-Cloud System and Big Data Engineering(Confluence). Noida, India, 2016.
22. Ellis D, Poliner G. Identifying cover songs' with chroma features and dynamic programming beat tracking[C]. ICASSP. Honolulu, Hawaii, USA, 2007.
23. Muller S E M., Kreuzer S. Making chroma features more robust to timbre changes[C].ICASSP. Taipei, Taiwan, 2009.
24. Richard M R G., Essid S. Combined supervised and unsupervised approaches forautomatic segmentation of radiophonic audio streams[C]. ICASSP. Honolulu, Hawaii,USA, 2007.