# Improving singing voice separation using attribute-aware deep network

Rupak Vignesh Swaminathan
*Alexa Speech*
*Amazon.com, Inc.*
United States
swarupak@amazon.com

Alexander Lerch
*Center for Music Technology*
*Georgia Institute of Technology*
United States
alexander.lerch@gatech.edu

***Abstract*—Singing Voice Separation (SVS) attempts to separate the predominant singing voice from a polyphonic musical mixture. In this paper, we investigate the effect of introducing attribute-specific information, namely, the frame level vocal activity information as an augmented feature input to a Deep Neural Network performing the separation. Our study considers two types of inputs, i.e, a ground-truth based 'oracle' input and labels extracted by a state-of-the-art model for singing voice activity detection in polyphonic music. We show that the separation network informed of vocal activity learns to differentiate between vocal and non-vocal regions. Such a network thus reduces interference and artifacts better compared to the network agnostic to this side information. Results on the MIR1K dataset show that informing the separation network of vocal activity improves the separation results consistently across all the measures used to evaluate the separation quality.**

***Index Terms*—Singing Voice Separation, Vocal Activity Detection, Deep Neural Networks, Attribute-aware training.**

## I. Introduction

Blind Audio Source Separation (BASS) is a widely explored topic by researchers in the audio processing field, especially Automatic Speech Recognition (ASR) and Music Information Retrieval (MIR). BASS plays an important role in ASR/MIR systems, as audio signals are mixtures of several audio sources (for example: background noise interfered with speech signals, multiple musical instruments playing at the same time) with little information about the sources. Usually, a pre-processing stage separates the sources, which often improves the accuracy of ASR/MIR systems [1], [2]. A well-known problem in the family of BASS is Singing Voice Separation (SVS), which is the task of isolating predominant vocals from a polyphonic musical mixture. SVS finds a wide variety of applications and serves as a pre-processing step in MIR tasks such as removal of vocals in karaoke systems, lyrics-to-audio alignment, singer recognition and main melody extraction [3]–[7].

Owing to its applications, the relevance of SVS has grown extensively in the last few years with several research groups contributing novel methods, datasets and evaluation metrics which are well documented as a part of the Signal Separation Evaluation Campaign (SiSEC) [8], [9]. Although the performance of SVS systems has improved over the last decade, the results show that there is still considerable room for improvement.

In this paper, we analyze how a neural network with a standard architecture for SVS can yield improved performance if its input feature set is augmented with vocal activity information. The vocal activity information, i.e., the indication that whether a frame contains vocals or not, is fed to the network as a one-hot encoded vector in addition to the Short-Time Fourier transform (STFT) magnitude of the polyphonic mixture. The research question we would like to address is whether this additional input can improve the system performance and how the system is impacted by the errors in vocal activity input. The main contribution of this paper is the systematic evaluation of an SVS network augmented with vocal activity information in order to improve the separation performance of the SVS network. We also quantify the effect of vocal activity in SVS by randomly perturbing the labels, injecting errors into the separation network and analyzing its performance.

The remainder of the paper is organized as follows. Section II discusses previous work done in SVS, informed source separation, and singing voice detection. Section III introduces our methodology. Section IV describes the experimental setup and the dataset. The results are presented and discussed in Section V. Finally, section VI summarizes our findings and presents directions for future work.

## II. Related Work

Successful approaches to the SVS task include techniques involving non-negative matrix factorization [10]–[12], probabilistic latent component analysis [13] and Bayesian adaptation methods [14]. Prior to the recent surge of deep learning models, techniques such as REpeating Pattern Extraction Technique (REPET) [15] and Robust Principal Component Analysis (RPCA) [16] had gained popularity for exploiting repeating patterns over a non-repeating melody (for example: repeating chord progressions and drum loops over lead vocals).

One of the earliest neural network models for this task was proposed by Huang et al. [17] in which a Deep Recurrent Neural Network (DRNN) architecture, having full temporal connections with a discriminative training procedure, predicted separate STFT magnitude targets for vocals and accompaniment.

This work was done prior to joining Amazon.com, Inc. while the first author was a graduate student at Georgia Institute of Technology.

Roma et al. used a DNN to estimate a time-frequency mask which is refined using F0 estimation to yield better performance [18]. A recent work by Uhlich et al. improved the state-of-the-art SVS results using by data augmentation and network blending with Wiener filter post processing [19].

Recently, several novel network architectures borrowed from related fields such as speech recognition, computer vision and biomedical signal processing have been successfully applied to this task. A convolutional encoder-decoder architecture that learns a compressed representation in the encoding stage and performs deconvolution during decoding stage to separate vocal and accompaniment was proposed in [20]. Deep U-net architecture, which was initially developed for medical imaging, was applied to SVS by Jansson et. al. [21] and was built on top of the convolutional encoder-decoder architecture while addressing the issue of lost details during encoding.

Attribute aware training, better known as informed source separation in the context of SVS has been an active area of research lately [22]–[26]. Although some techniques for score-informed musical source separation have been proposed in [22], [26], the availability of scores may pose problems [25]. Attribute aware training has been well-studied in speech recognition [27]–[29] where separately trained acoustic embeddings or speaker derived i-vectors [30] have been used to augment the input feature set to improve the results on speech recognition. A closely related work used a two stage DNN architecture for speech denoising in low SNR environments [31]. The output of a speech activity detection network was fed into a denoising autoencoder, enabling better speech denoising with the implicitly computed noise statistics.

Vocal activity-informed RPCA was one of the earlier works to incorporate vocal activity information in the RPCA framework for SVS [24]. It was shown that the vocal activity-informed RPCA algorithm outperformed the system uninformed of vocal activity. In this work, we use the state-of-the-art singing voice detection model proposed in [32] to improve the performance of the SVS network and compare it to the network agnostic to the additional attribute information.

## III. System

Figure 1 shows the overall structure of the system being evaluated. The SVS system is being fed additional input about vocal activity. The output of the network is the estimated magnitude spectra of the vocals and accompaniment which are inverted using the phase of the input polyphonic mixture.

### A. Singing Voice Separation Network

Our model for SVS is a simple multi-layer feedforward neural network with separate targets for vocals and accompaniment [17]. Our system is a 3-layer feedforward neural network with 1024 hidden neurons each, and the input representation is a STFT magnitude of the polyphonic mixture. The STFT is extracted with a 1024-point FFT, frame size of 640 samples and hop size of 320 samples (audio clips sampled at 16KHz). Additionally, it is stacked with neighbouring audio frames as suggested in [17] to add contextual information resulting in an dimensionality of $3 \cdot 512$. The targets are STFT magnitude of the separated vocals and accompaniment.

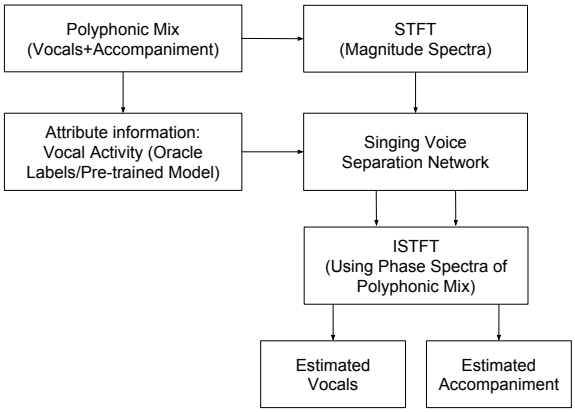


Fig. 1. Block diagram of Singing Voice Separation network informed of vocal activity. The network predicts STFT magnitude of the sources (vocals and accompaniment) which are combined with the STFT phase of the input polyphonic mixture to reconstruct the waveforms of the respective sources.

We train this network with a joint mask training procedure as proposed in [33]. According to this procedure, the outputs of the penultimate layer ($\widehat{y}_1$ and $\widehat{y}_2$) of the separation are used to compute a soft time-frequency mask. The targets of the separation network, $\widetilde{y}_1$ and $\widetilde{y}_2$ are estimated by taking the Hadamard product between the result of soft time-frequency masking layer and the input magnitude spectra of the polyphonic mixture (denoted by z).

$$\widetilde{y}_1 = \frac{|\widehat{y}_1|}{|\widehat{y}_1| + |\widehat{y}_2|} \odot z \tag{1}$$

$$\widetilde{y}_2 = \frac{|\widehat{y}_2|}{|\widehat{y}_1| + |\widehat{y}_2|} \odot z \tag{2}$$

The objective function to train the network is the sum of the mean squared error between the network predictions ($\widetilde{y}_1$, $\widetilde{y}_2$) and the clean sources ($y_1$, $y_2$).

$$J = \|\widetilde{y}_1 - y_1\|_2^2 + \|\widetilde{y}_2 - y_2\|_2^2 \tag{3}$$

The outputs of separation network, $\widetilde{y}_1$ and $\widetilde{y}_2$, are combined with the phase spectra of the original polyphonic mixture to obtain complex spectra. We use overlap and add method to reconstruct the respective vocal and accompaniment waveforms.

### B. Vocal Activity Information

*1) Oracle Labels:* We present the ground truth frame level vocal activity along with the magnitude spectrum of the input polyphonic mixture to the SVS network to observe its separation quality. The labels are represented as a one-hot encoded vector of two dimensions. This is considered the best case scenario where the labels are known during training and inference. To further evaluate the performance under real-world scenario, we use a model for vocal activity detection during inference which is described below.

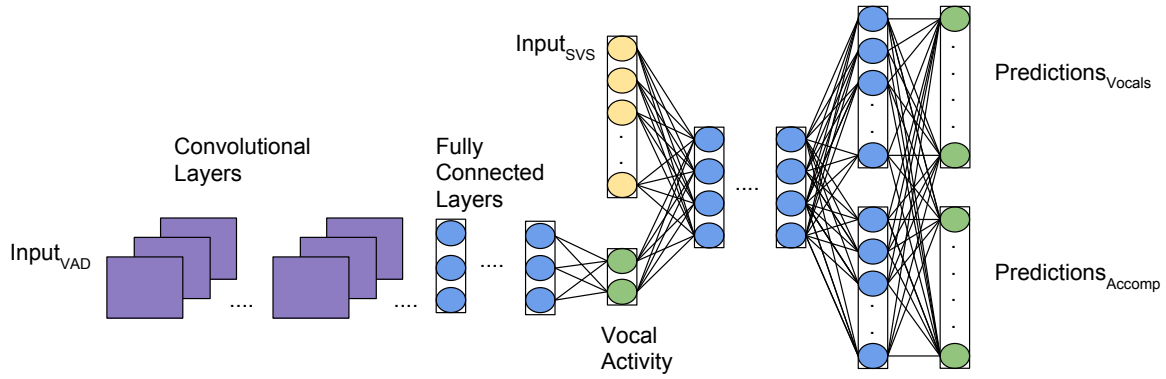


Fig. 2. Modular DNN framework consisting of a CNN-based Vocal Activity Detection network and a multi-layered feed forward Singing Voice Separation network. Input$_{VAD}$ is log-mel spectrogram with 20 context frames on either side and Input$_{SVS}$ is magnitude spectrogram of the mixture with a single frame of context on either side. Predictions$_{Vocals}$ and Predictions$_{Accomp}$ are the estimated magnitude spectra of the separated sources.

*2) Vocal Activity Detection Model:* Vocal Activity Detection (VAD) or Singing Voice Detection is closely related to timbre classification/instrument recognition. Therefore a number of previous works follow similar approaches of classifying segments/frames by learning timbre. It has been shown that with a long context logmel input representation, Convolutional Neural Networks (CNNs) outperform most of the other architectures [32]. Hence, we use CNNs for learning singing voice characteristics and train it to output vocal activity predictions which are fed into the SVS network as shown in Figure 2. The network has the following architecture:

(i) A convolutional layer with 64 features maps and a 3x3 kernel,
(ii) A 2x2 maxpooling layer,
(iii) A convolutional layer with 32 feature maps and a 3x3 kernel,
(iv) A 2x2 maxpooling layer,
(v) 2 convolutional layers with 128 and 64 features maps each with 3x3 kernels,
(vi) 2 dense layers of size 512 and 128,
(vii) An output layer of size 2.

The hidden layers have Relu non-linearity and the output layer has a softmax activation. The input representation is log-mel spectrogram with 80 filterbanks and 40 neighbouring context frames (20 on either side of the center frame) with the voicing label corresponding to the center frame. The model is trained with a cross-entropy loss between the targets and the one-hot encoded labels, optimized with Adadelta optimizer. The architecture is a slightly modified version of the state-of-the-art singing voice detection algorithm presented in [32].

## IV. Experimental setup

### A. Dataset

We use the MIR1K dataset throughout our experiments [34]. The dataset contains 1000 snippets (total of 133 minutes) of Chinese karaoke performances sampled at 16 kHz. It has vocals and accompaniment tracks separated in two channels. The vocal activity labels are annotated at the frame level with a frame size of 40 ms and hop size of 20 ms. The data split (Train/Test/Validation) is the same as in [17].

### B. Methodology

We investigate the following scenarios during training and inference of the SVS network:

- Case 0: No vocal activity information,
- Case I.a: Using oracle vocal activity labels (ground truth) during training and inference,
- Case I.b: Perturbing the oracle vocal activity labels by injecting errors at various error percentage levels during training and inference, and
- Case II: Using a pre-trained model for vocal activity detection during inference to evaluate a real-world use case. The output predictions (softmax probabilities) are fed into the separation network as shown in Figure 2.

### C. Evaluation Metrics

To evaluate the quality of separation, standard performance measures for blind source separation of audio signals (BSS Eval measures) [35] are used. These metrics include Source-to-Distortion Ratio (SDR), Source-to-Artifacts Ratio (SAR), and Source-to-Interference Ratio (SIR). The estimated signal is decomposed into target distortion, interference, and artifacts which are used to compute the scores. The estimated signal having minimal distortion, interference, and artifacts, will result in high scores. A Normalized SDR measure is computed as defined in [17] and global scores (GNSDR, GSAR and GSIR) are reported. The global scores are computed by taking the weighted average of the individual scores of the audio files, weighted by their length.

### D. Model Selection and Generalization

To prevent overfitting, the training in both SVS and VAD is stopped as early as the validation loss starts to increase, and the hyperparameters are selected based on the vocal

| | Predicted: No-Vocal | Predicted: Vocal |
|---|---|---|
| True: No-Vocal | 60659 (85.08%) | 10635 (14.92%) |
| True: Vocal | 10732 (4.18%) | 246269 (95.82%) |

TABLE I
CONFUSION MATRIX FOR THE CNN VOCAL ACTIVITY DETECTION MODEL.

| Model | GNSDR | GSAR | GSIR |
|---|---|---|---|
| Without DA | 6.04 | 8.77 | 10.88 |
| With DA | 6.77 | 9.52 | 11.45 |

TABLE II
EFFECT OF DATA AUGMENTATION

| Model | GNSDR | GSAR | GSIR |
|---|---|---|---|
| Case 0 | 6.77 | 9.52 | 11.45 |
| Case I.a | 7.16 | 9.86 | 11.72 |

TABLE III
USING CLEAN ORACLE LABELS DURING TRAINING AND INFERENCE

| Perturb (%) | GNSDR | GSAR | GSIR |
|---|---|---|---|
| 0 | 7.16 | 9.86 | 11.72 |
| 2.5 | 6.90† | 9.90 | 11.12 |
| 5 | 6.95 | 9.75 | 11.45 |
| 7.5 | 6.86† | 9.84 | 11.13 |
| 10 | 6.73 | 9.71 | 11.02 |
| 15 | 6.69 | 9.69 | 10.94 |

TABLE IV
SEPARATION RESULTS FOR TRAINING AND INFERENCE WITH PERTURBED VOCAL ACTIVITY LABELS. STATISTICALLY INSIGNIFICANT RESULTS ARE DENOTED BY †.

| Perturb (%) | GNSDR | GSAR | GSIR |
|---|---|---|---|
| 0 | 6.99 | 9.72 | 11.54 |
| 2.5 | 6.97 | 9.93 | 11.24 |
| 5 | 6.93 | 9.73 | 11.43 |
| 7.5 | 6.90 | 9.85 | 11.21 |
| 10 | 6.74 | 9.71‡ | 11.07 |
| 15 | 6.72 | 9.71‡ | 11.01 |

TABLE V
SEPARATION RESULTS FOR TRAINING WITH PERTURBED VOCAL ACTIVITY LABELS AND INFERENCE USING CNN VOCAL ACTIVITY DETECTION MODEL. STATISTICALLY INSIGNIFICANT RESULTS ARE DENOTED BY ‡.

GNSDR results on the validation set. It should be noted that the amount of the training data (171 audio clips) is quite small compared to the test set (825 audio clips), which a reason for concern when training DNNs. As a generalization strategy to overcome the problem of overfitting, we train the separation network by randomly shuffling the accompaniment every epoch before mixing them with the vocals at the input of the separation network. This Data Augmentation (DA) procedure virtually increases the number of training examples and helps the separation network perform better on unseen examples. Previous works [17], [19] have proposed similar DA strategies to prevent overfitting.

## V. RESULTS AND DISCUSSION

### A. Vocal Activity Detection

Before we start our planned experiment, the performance of the CNN-based Vocal Activity Detection model has to be determined on the test set of MIR1K. The confusion matrix is shown in Table I. It is observed that the model performs reasonably well with an accuracy of 93.5% and F1 score of 0.95. This is consistent with the results reported with a similar architecture on standard singing voice detection datasets [32].

### B. Data Augmentation for Singing Voice Separation

Table II shows the effect of training with random shuffling of accompaniment in every epoch. It is observed that DA indeed improves the performance of the model. We will use this data augmented model throughout the rest of our experiments.

### C. Case 0 and Case I.a: Impact of Oracle Labels

To confirm our hypothesis that the vocal activity information helps the separation network learn better while reducing artifacts and interference, we model a best case scenario by feeding the ground truth labels from the dataset to the SVS network. The results of using clean oracle labels during training and inference of the separation network is shown in Table III.

### D. Case I.b: Perturbed Oracle Labels

The results of separation network augmented with perturbed oracle labels are shown in Tables IV. It can be observed that as we increase the perturbation, the separation quality drops proportionally. It is interesting to note that training with perturbation beyond 10% makes the separation network perform at par or slightly worse than the network not informed of vocal activity. This elucidates the sensitivity of the separation network to the vocal activity labels.

### E. Case II: Using pre-trained vocal activity during inference

Finally, we report the results of CNN vocal activity detection model during inference (Table V) The separation network behaves in the same manner as in the previous case as the separation performance decreases with increase in perturbation.

### F. Discussion

To measure the significance of our results, we perform pair-wise t-tests to confirm whether (a) Vocal activity informed SVS is better than the network uninformed of vocal activity and (b) As the perturbation increases, the separation quality decreases. We confirm that all our results are statistically significantly with $p < 0.05$, except the pairs denoted by † and ‡ in Table IV and V, respectively.

It can be observed from Table II that DA improves the separation results significantly and consistently across all three measures. In the best case scenario of feeding unperturbed oracle labels during inference, we observe the best separation results which confirms our hypothesis that vocal activity information helps in better separation performance of the DNN. It is interesting to note that vocal activity informed RPCA [24] did not show any improvements on GSAR while our vocal activity informed DNN shows consistent improvements across all three evaluation measures.

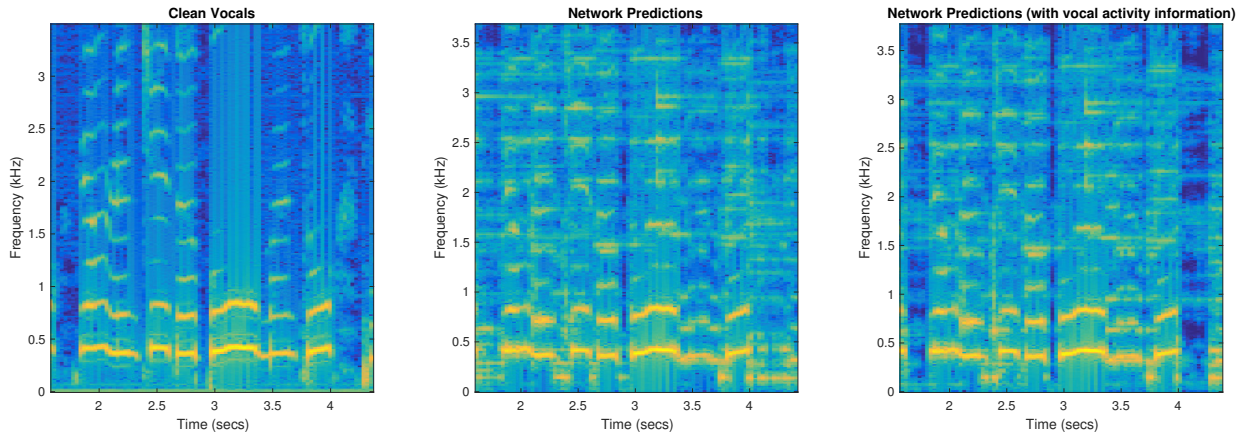


Fig. 3. Spectrograms of Clean Vocals, Network Predictions in Case 0, and Network Predictions in Case I.a. Observe the interference and artifacts present in Case 0, especially in unvoiced regions, and how they are improved when vocal activity information is considered (Case I.a).

In order to investigate what the network learns when augmented with vocal activity, we plot the (a) Spectrograms of clean vocals, (b) Network predictions of Case 0 and (c) Network predictions of Case I.a. It can be inferred from the spectrograms that for non-vocal regions, the artifacts and interference are much lesser for Case I.a compared to Case 0, suggesting that the network learns to differentiate between vocal and non-vocal regions and suppress regions in the polyphonic mixture that do not contain vocals and emphasize on the regions with vocal activity. We also plot the saliency map of the network [36] which is defined as the derivative of the output of the network with respect to the input, in order to understand how the trained network forms its decisions. From saliency maps, we can infer which parts in the input are most crucial to the network and influence the output of the network.

It can be seen from Figure 4 that the saliency map of the vocal activity informed network reveals more characteristics of singing voice (better harmonic structure) compared to the case without vocal activity. Also, it can be observed once again that the network looks only at the vocal portions of the input and the non vocal portions are set to almost zero whereas in the case of the network agnostic of vocal activity does not differentiate between vocal frames vs. non-vocal frames.

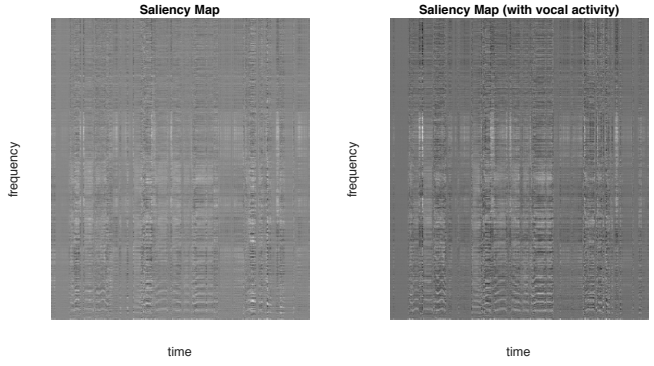


Fig. 4. Saliency maps

In Case I.b we ascertain the susceptibility of the separation network to perturbed oracle labels. As expected, the separation performance decreases consistently as the perturbation is increased. Case II emulates a real-world scenario where the vocal activity labels are unknown during inference and a model is needed to predict the vocal activity. In comparison to the Case I.b of testing with perturbed oracle labels, this inference with CNN VAD model is slightly better. Our conjecture is that since the distributions of perturbations are quite different in these two cases, the separation network might not regard the errors in case of random perturbations equally as the errors made by the CNN VAD model. Since the random perturbations are drawn from an uniform distribution, it corrupts "easy" examples for the separation network as equally likely as the "hard" examples. Therefore, in the case where the random perturbation corrupts an easy example, the separation network outputs poor predictions which would have been otherwise predicted easily. On the other hand, we believe that the examples that are hard to learn for the CNN VAD model are the outliers which are hard to learn even for the separation network. Hence, the predictions of the separation network for "easy" examples are always going to be better when the vocal activity labels from the CNN VAD model are used instead of perturbed oracle labels.

## VI. Conclusion and Future Work

We studied the effect of augmenting the separation network with vocal activity labels during training and testing of a DNN performing SVS. The vocal activity labels are either ground truth labels, distorted ground truth labels, or labels predicted with a state-of-the-art CNN VAD model. We showed that the separation network is able to learn about the regions of vocal activity and reduces artifacts and interference in the non-vocal regions. As a future direction of this research, we would like to explore more attributes that could be fed as additional inputs, such as singer-specific features (i-vectors) and lyric-specific features (lyric-audio alignment) so as to improve SVS.

## REFERENCES


[1] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[2] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 169–172.

[3] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2007, pp. 375–378.

[4] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1475–1487, 2007.

[5] S. W. Lee and J. Scott, "Word level lyrics-audio synchronization using separated vocals," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 646–650.

[6] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, p. 4, 2010.

[7] Y. Ikemiya, K. Itoyama, and K. Yoshii, "Singing voice separation and vocal f0 estimation based on mutual combination of robust principal component analysis and subharmonic summation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2084–2095, 2016.

[8] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 323–332.

[9] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 293–305.

[10] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2004, pp. 494–499.

[11] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*. Citeseer, 2005, pp. 337–344.

[12] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[13] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separating a foreground singer from background music," in *Proc. Int. Symp. Frontiers Res. Speech Music.*, 2007.

[14] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1564–1578, 2007.

[15] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (repet): A simple method for music/voice separation," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 1, pp. 73–84, 2013.

[16] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 57–60.

[19] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 261–265.

[17] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 477–482.

[18] G. Roma, E. M. Grais, A. J. Simpson, and M. D. Plumbley, "Singing voice separation using deep neural networks and f0 estimation."

[20] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 258–266.

[21] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 323–332, 2017.

[22] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.

[23] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*. IEEE, 2013, pp. 1–4.

[24] T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and R. Jang, "Vocal activity informed singing voice separation with the ikala dataset," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 718–722.

[25] S. Ewert, B. Pardo, M. Müller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[26] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 888–891.

[27] J. Rownicka, P. Bell, and S. Renals, "Analyzing deep cnn-based utterance embeddings for acoustic model adaptation," *arXiv preprint arXiv:1811.04708*, 2018.

[28] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[29] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 225–229.

[30] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors." in *ASRU*, 2013, pp. 55–59.

[31] P. G. Shivakumar and P. G. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement." in *INTERSPEECH*, 2016, pp. 3743–3747.

[32] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks."

[33] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1562–1566.

[34] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[35] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[36] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.