Interspeech 2018
2-6 September 2018, Hyderabad

Music Source Activity Detection and Separation using Deep Attractor Network

Rajath Kumar *Yi Luo* *Nima Mesgarani*

Department of Electrical Engineering, Columbia University, New York, NY

rm3497@columbia.edu yl3364@columbia.edu nima@ee.columbia.edu

Abstract

In music signal processing, singing voice detection and music source separation are widely researched topics. Recent progress in deep neural network based source separation has advanced the state of the performance in the problem of vocal and instrument separation, while the problem of joint source activity detection and separation remains unexplored. In this paper, we propose an approach to perform source activity detection using the high-dimensional embedding generated by Deep Attractor Network (DANet) when trained for music source separation. By defining both tasks together, DANet is able to dynamically estimate the number of outputs depending on the active sources. We propose an Expectation-Maximization (EM) training paradigm for DANet which further improves the separation performance of the original DANet. Experiments show that our network achieves higher source separation and comparable source activity detection against a baseline system.

Index Terms: Source Activity Detection, Source Separation, Deep learning, Deep Attractor Network

1. Introduction

The increasing popularity of cloud based music subscription services, requires more advanced classification/recommendation systems that cater to the interest of each individuals. As these involve large catalogues of music tracks across different countries and artists, there is a need for advanced indexing based on the sources (vocal/instruments) utilized in a track. In this paper, we propose a new front-end system that localizes each source present in a music track. This proposed front-end framework can be applied to varied applications ranging from query-by-singer, query-by-instruments, query-by-lyrics to recommendation systems based on similar rhythmic structure of instruments. We mainly try to answer the question "*which vocal/instrument played when?*".

Two main activity detection systems related to our work include voice activity detection (VAD) and singing voice detection (SVD). Both systems distinguish vocal/speech segments from the rest. Conventional approaches in SVD involve traditional machine learning algorithms [1, 2], and energy-based systems [3, 4] are widely used for VAD. In recent years, neural network based systems have been applied to activity detection and have shown to be able to generalize well [5, 6, 7, 8]. Since then, much of the research in activity detection has largely focused on engineering robust features for deep learning based approaches [9, 10, 11, 12].

In general, source separation problems can be categorized into two classes. *In-class separation* aims at separating sources that are similar or in the same category, such as speech versus speech or similarly pitched instruments. *Between-class separation* corresponds to the separation of different type of sources, such as vocal/accompaniment separation. Most of the traditional methods like computational auditory scene analysis [13], robust principal component analysis [14] and low-rank modeling [15] may perform well in between-class separation tasks, but are not robust enough for in-class separation problems. With the development of deep learning, neural networks for time-frequency (T-F) mask inference [16, 17] greatly improved the performance and robustness in the between-class tasks. The general paradigm in those systems is to perform short-time analysis on the audio mixture. The mixture spectrogram is fed as the input to the network, and an estimated T-F mask for each source is considered as the output. Although those networks have been successful in music separation, they typically rely on the effectiveness of independent modeling for each source [16, 18, 19]. These independent networks are not reasonable for scaling the separation task to many sources. To reduce the computational complexity involved in independent modeling, there has been several attempts at sharing feature/layer across the target sources [17, 20]. But as we will show by experiments, these models have limited success in in-class separation. Also during the training of those models, it is often assumed that all the sources are active in the mixture, which is not the case in real-world scenario.

To overcome the in-class separation limitation in feature/layer sharing network and to stably scale the number of sources, we propose an end-to-end clustering based approach, Deep Attractor Network (DANet) [21]. Clustering based approaches for mask estimation have become the state-of-the-art in speaker separation since the introduction of deep clustering [22]. Following the idea, ChimeraNet [20] utilized both the deep clustering and mask inference methods in a multi-task fashion and achieved remarkable improvements over conventional approaches in between-class music separation.

We investigated both the ChimeraNet and DANet on separation and activity detection performance. We also discuss an implicit learning method of source activity in the high dimensional embedding space of DANet. Upon analyzing this embedding space, we find that the clusters of similarly pitched instruments lie close to one another. To better estimate and model these clusters distinctively, we propose a complete Expectation-Maximization training of DANet which significantly improves the performance of in-class separation. The effectiveness of DANet in music separation and activity detection demonstrates that clustering based approaches are competitive alternative to direct mask-inference using feature/layer sharing networks that are widely applied in these tasks [17, 23, 20, 24].

2. Music Source Separation

We model our system in time-frequency (T-F) domain [20, 21, 25]. The single-channel music mixture $x(t) \in \mathbb{R}^{1\times T}$ is the sum of its C sources $s_1(t), s_2(t), ...s_c(t)$. Thus in the complex T-F domain, the short-time Fourier transform (STFT) of the mixture $X(f,t)$ is the sum total of STFT's of its sources $S_{i=1:C}(f,t)$. The flattened magnitude spectrogram of the mixture $|X_{ft}| \in \mathbb{R}^{1\times FT}$ is fed as input to the network. We employ

10.21437/Interspeech.2018-2326

T-F masking technique to estimate the individual sources. In particular, we use wiener-filter like mask (WFM) as the oracle source assignment represented as $m_i \in \mathbb{R}^{1\times FT}$. The WFM, m_i is computed as follows,

$$m_{i,ft} = \frac{|S_{i,ft}|^2}{\sum_{i=1}^{C}|S_{i,ft}|^2}, \qquad s.t. \quad \sum_{i=1}^{C} m_i = 1 \tag{1}$$

The spectrograms of the individual sources are estimated by performing element-wise multiplication between the mixture magnitude spectrogram and the masks

$$|\hat{S}_{i,ft}| = |X_{i,ft}| \odot m_i \tag{2}$$

Further, inverse STFT is applied to obtain the reconstructed waveform $\hat{s}_1(t), \hat{s}_2(t), ...\hat{s}_c(t)$ using the estimated magnitude spectrograms $|\hat{S}_i|$ and the phase information $\angle X(t,f)$ of the mixture.

2.1. Deep Attractor Network

The separation problem in Deep Attractor Network (DANet) is formulated as a multi-class regression problem in a supervised setting. For each of the T-F bin, a K dimensional embedding, $V \in R^{K\times FT}$ is generated through Bi-directional Long Short Term Memory (Bi-LSTM) layers. The attractors $a_{i=1:C} \in R^{1\times K}$ are calculated for mask estimation using only the prominent T-F bins of the mixture. These prominent bins are determined by a binary vector, $\mathbf{w} \in R^{1\times FT}$ obtained by thresholding the power in each of the spectrogram.

$$a_i = \frac{(\mathbf{y}_i \odot \mathbf{w})V^T}{\sum_{f,t}(\mathbf{y}_i \odot \mathbf{w})}, \qquad i = 1, 2, ...C \tag{3}$$

$\mathbf{y}_{i=C} \in R^{1\times FT}$ is the known source assignment of each T-F bin. The masks are then estimated by calculating the distance between the attractors and the embeddings.

$$\hat{m}_{i=1:C} = Softmax(a_{i=1:C}V) \tag{4}$$

Softmax activation is applied to satisfy equation 1. Using the evaluated masks, the objective function of the network is the standard mean squared error

$$l = \frac{1}{C}\sum_i ||X \odot (m_i - \hat{m}_i)||_2^2 \tag{5}$$

2.2. Anchored Deep Attractor Network

The knowledge of the source assignment $\mathbf{y}_i$ of each T-F bin was necessary to compute the attractors in DANet [21]. However in Anchored Deep Attractor Network (ADANet) [25] the source assignment labels are not required. These assignments are estimated with the help of reference points/anchors in the embedding subspace, $R^{K\times FT}$.

During training, we initialize N random trainable anchor points, $b_{j=1:N} \in R^{1\times K}$. For music separation, N is equal to the number of sources C as the sources to be distinguished are known. We also don't perform any permutation of the anchor points as done in [25]. This ensures that each anchor point initialized corresponds to a single source. The distance between the anchors and the embeddings is used to compute the source assignment, $\hat{\mathbf{y}}_{i=1:N} \in R^{C\times FT}$ (note that $N = C$). This computation is similar to equation 4. Using these estimated assignments, attractors are calculated using equation 3.

2.3. Expectation Maximization Training of DANet

Individual sources in music tracks exhibit high temporal correlation. Most of the cases, multiple instruments play the same note and are tuned to the same frequency scale. When we model this data for separation using DANet, there is a high probability that the clusters overlap. This leads to attractors being formed close to each other resulting in poor mask estimation. To circumvent this issue, we consider the variance of the overlapping clustered data as in Gaussian mixture models to compute the masks. Thus we propose an Expectation Maximization (EM) [26] framework for training DANet.

Using the assignments generated by the anchors as discussed in Section 2.2, G set of Gaussian component probability density functions, $\mathcal{N}_{j=1:G} \triangleq \mathcal{N}(x|\mu_{j=1:G}, \Sigma_{j=1:G})$ (note that $G = C$) is initialized. During the E-step, the probability that K dimensional embeddings $V_{i=1:FT} \in R^{K\times FT}$ belongs to the Gaussian component $\mathcal{N}_j$ is estimated. This posterior probability, $P(N_j/V_i)$ is the estimated mask. E-step is described as follows,

$$P(\mathcal{N}_{j=1:G}|V_{i=1:FT}) = \frac{P(V_i|\mathcal{N}_j)P(\mathcal{N}_j)}{P(V_i)} \tag{6}$$

Consider mixing coefficient α, then equation 6 can be reformulated as,

$$P(\mathcal{N}_j|V_i) = \frac{\alpha_j\mathcal{N}(V_i|\mu_j,\Sigma_j)}{\sum_{j=1}^{G}\alpha_j\mathcal{N}(V_i|\mu_j,\Sigma_j)} \tag{7}$$

During the maximization step, we recompute the Gaussian component parameters using the equations described below,

$$\begin{aligned} \mu_j &= \frac{\sum_i^{FT} P(\mathcal{N}_j|V_i)V_i}{\sum_i^{FT} P(\mathcal{N}_j|V_i)} \\ \Sigma_j &= \frac{\sum_i^{FT} P(\mathcal{N}_j|V_i)(V_i-\mu_j)(V_i-\mu_j)^T}{\sum_i^{FT} P(\mathcal{N}_j|V_i)} \\ \alpha_j &= \frac{\sum_i^{FT} P(\mathcal{N}_j|V_i)}{FT} \end{aligned} \tag{8}$$

The EM procedure is executed only for a single step for every iteration. The objective function remains the same as equation 5.

3. Source Activity Detection

3.1. Activity detection using DANet

Activity detection is performed using the trained separation network described in Section 2. Note that the network was never given any information on silence or active sources. The reference points/anchors used by ADANet and EM-DANet provides a cue on the location of embedding in its own subspace. Thus, we are effectively making use of the location of embeddings generated by a classification network trained for separation, to achieve learning for activity detection task. Methods employing similar fashion include time contrastive learning [27] and generative adversarial networks [28] where the network is trained to classify on one task, but the learned embeddings were robust for other tasks.

The distance between the T-F embedding and the anchors localizes the points in the sub-space. This distance is the source assignment $\hat{\mathbf{y}}$ computed from equation 4. Figure 1(a) visualizes the same. If $\mathbf{q}$ is considered to be the number of bins in a T-F

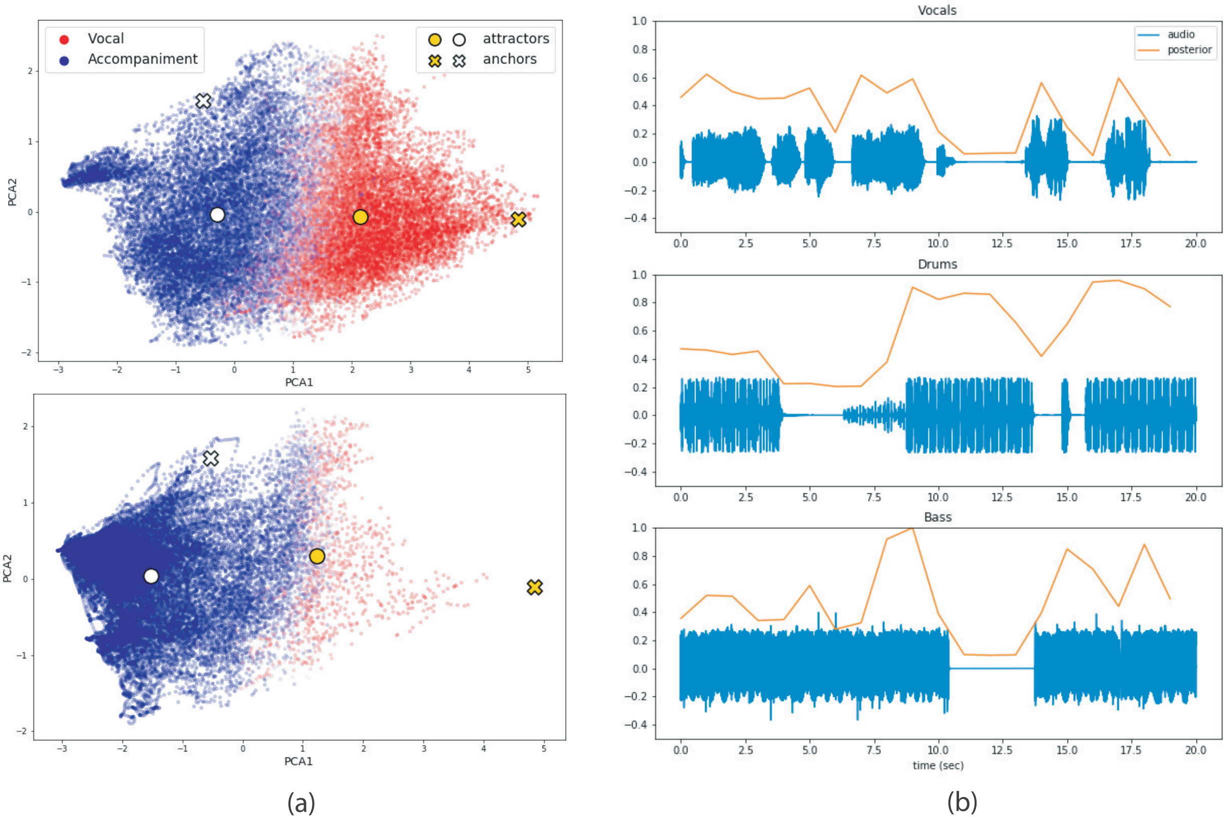

Figure 1: *Unsupervised source activity detection in deep attractor network. (a) Clustering of vocals and accompaniment visualized using their source assignment as magnitude along with their respective anchors and attractors for 2 separate scenarios. (top) both the sources are active; (bottom) only accompaniment is active. (b) The activity detection posteriors computed from the distance between the embedding and anchor point plotted for different sources of Music.*

segment greater than a particular distance threshold t then the posteriors, $\mathbf{p}_{i=1:C} \in R^{C\times T}$ for activity detection are generated as follows,

$$\mathbf{p}_{i=1:C} = \frac{\mathbf{q}_{i=1:C}}{T \times F} \tag{9}$$

Essentially the maximum posterior value is attained when all the T-F bins of a segment are closest to its anchor. This also ascertains that the source is active in that time period. During inference, having the knowledge of these posteriors for each source enables the network to dynamically output masks by thresholding the posterior value. We can also infer the major contributing sources in a particular segment by picking the top n posteriors or single out the major source by selecting the max posterior.

3.2. Activity detection in mask-inference networks

In a direct mask-inference network such as [20], active and inactive sources can be determined by investigating the salient T-F points in the estimated masks. Investigating the embedding space is not possible in these type of networks. Thus we rely on the assumption that the silent segments in a source channel consists of white noise below the energy levels of active sources [29]. The source assignment labels that are provided will reflect this information. In the case of ADANet and EM-DANet, this information is estimated using anchors without prior knowledge. Thus identifying the active output streams in ChimeraNet involves thresholding at lower values and the formulation is the same as equation 9.

4. Evaluation

4.1. Dataset

We make use of professionally mixed musdb18 dataset [30]. It consists of 150 tracks (~10 hours) split into 100 songs for training and 50 for test. To improve the performance of the system, the train dataset is augmented by remixing the sources. ~13.3 hours of train data is created along with a separate ~3.5 hours of development dataset. The single-channel audio is downsampled to 16 KHz.

The input feature is computed using STFT with 2048 and 512 point window-size and hop-size respectively. To reduce the computational complexity, a 300 dimensional mel-filterbank is multiplied to scale down the spectrogram. The features were segmented at every 100 frames (~3.2 seconds).

The posteriors for activity detection were evaluated for every segment. The true label to evaluate the detection was computed from an energy level based activity detector.

4.2. Network architecture

All the models tabulated contains 4 Bi-directional LSTM layers with 600 hidden units in each layer. 20 dimension embedding sub-space is used similar to [21, 25]. Dropout probability is set

		nSDR	SIR	SAR
Chimera	Vocal	**8.99**	**12.02**	2.95
	Other	4.58	6.43	0.32
	Bass	6.82	**8.38**	2.06
	Drums	6.99	11.59	4.10
DANet	Vocal	8.14	7.78	2.84
	Other	4.34	2.98	0.57
	Bass	5.53	2.26	**3.29**
	Drums	4.96	8.44	2.42
ADANet	Vocal	7.31	10.06	1.34
	Other	3.35	4.37	**1.29**
	Bass	6.26	5.84	2.02
	Drums	5.61	**14.68**	2.17
EM-DANet	Vocal	8.98	11.86	**2.98**
	Other	**4.59**	**6.63**	0.35
	Bass	**6.98**	8.09	2.20
	Drums	**7.28**	11.70	**4.40**

Table 1: *Music source separation results on 4 source condition*

		nSDR	SIR	SAR
Chimera	Vocal	9.73	16.51	3.30
	Accompaniment	4.26	18.57	12.09
DANet	Vocal	**9.79**	14.23	**3.56**
	Accompaniment	**4.38**	17.88	**12.39**
ADANet	Vocal	9.36	**16.59**	3.37
	Accompaniment	3.14	**19.51**	10.81
EM-DANet	Vocal	9.29	16.37	2.83
	Accompaniment	4.37	19.29	12.07

Table 2: *Music source separation results on 2 source condition*

to 0.5 in all the LSTM layers. Adam optimizer with learning rate $1e^{-4}$ is utlized to train the network. Learning rate is halved if the validation loss does not decrease in 3 epochs and all the networks are trained for 100 epochs.

4.3. Results

Music source separation has been evaluated using 3 parameters: signal-to-distortion ratio (SDR), signal-to-artifacts ratio (SAR) and signal-to-interference ratio (SIR). The separation results are shown in Table 1 and 2. The musdb18 dataset provides data of 4 sources for each track - vocal, bass, drums and others. The experiments were carried out to measure the performance of the network in between-class and in-class cases and thus 4 sources and 2 sources (accompaniment comprises of bass, drums and others) are considered.

Analysing 4 source condition, EM-DANet significantly outperforms ChimeraNet in in-class separation. EM-DANet also improves over the performance of existing DANet architectures. Both ChimeraNet and EM-DANet perform similarly in between-class separation of sources namely vocals. In 2 source condition all the networks output comparable performances. However DANet architectures perform better than ChimeraNet in all the considered metrics. It is evident from the results that as we scale the number of sources, clustering methods with the right optimization and training paradigms can outperform direct mask-inference methods.

		EER	AUC
ADANet	Vocal	23.38	84.10
	Drums	23.57	84.47
	Bass	34.89	70.46
EM-DANet	Vocal	**15.05**	89.65
	Drums	**18.13**	87.41
	Bass	25.41	**80.64**
Chimera	Vocal	16.25	**90.89**
	Drums	19.11	**88.26**
	Bass	**24.98**	79.94

Table 3: *Equal Error Rate (EER) and Area Under Curve (AUC) evaluated for activity detection task on 4 source condition.*

		EER	AUC
ADANet	Vocal	14.33	90.99
EM-DANet	Vocal	**14.11**	90.95
Chimera	Vocal	14.79	**92.01**

Table 4: *Equal Error Rate (EER) and Area Under Curve (AUC) evaluated for activity detection task on 2 source condition.*

To evaluate the activity detection performance, we plot a Receiver Operating Characteristic (ROC) curve. ROC curve was plotted against True Alarm Rate (TAR) and False Alarm Rate (FAR) to obtain Area Under Curve (AUC). Equal Error Rate (EER) is also shown and is defined as the point at which False Reject Rate (FRR) equals FAR. Better performing systems have a low EER and high AUC. The results of this task are shown in Table 4 and 3. Sources not representing a particular instrument were dropped from evaluation (other and accompaniments). Despite the difficulty to distinguish the silence segment of a source when several other sources are active in the same embedding sub-space, EM-DANet exhibits comparable performances with ChimeraNet across all the sources in both 2 source and 4 source condition. Amongst the DANet architectures, EM training of DANet provides remarkable improvements over previous training of ADANet. Figure 1(b) plots the posteriors generated by EM-DANet in 4-source scenario.

5. Conclusion

In this paper, we demonstrate the effectiveness of deep attractor network in the task of music source separation and activity detection. We have also proposed a representation learning approach to compute the activity detection by declaring an anchor in the embedding subspace. A detailed generalized Expectation-Maximization framework for training deep attractor network is discussed. From our experiments, this paradigm greatly benefits the separation of similar pitched/frequency instruments without any decline in performance of other metrics.

6. References

[1] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 1885–1888.

[2] W. Chou and L. Gu, "Robust singing detection in speech/music discriminator design," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 865–868.

[3] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 4, pp. 777–785, 1981.

[4] G. Evangelopoulos and P. Maragos, "Speech event detection using multiband modulation energy," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[5] G. Gelly and J.-L. Gauvain, "Minimum word error training of rnn-based voice activity detection," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7378–7382.

[7] R. Zazo Candil, T. N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform cldnns for voice activity detection," 2016.

[8] J. Schlüter, "Learning to pinpoint singing voice from weakly labeled examples." in *ISMIR*, 2016, pp. 44–50.

[9] B. Lehner, R. Sonnleitner, and G. Widmer, "Towards light-weight, real-time-capable singing voice detection." in *ISMIR*, 2013, pp. 53–58.

[10] A. L. Berenzweig, D. P. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society, 2002.

[11] B. Lehner, G. Widmer, and R. Sonnleitner, "On the reduction of false positives in singing voice detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7480–7484.

[12] L. Regnier and G. Peeters, "Singing voice detection in music tracks using direct voice vibrato detection," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 1685–1688.

[13] 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.

[14] 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.

[15] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Real-time online singing voice separation from monaural recordings using robust low-rank modeling." in *ISMIR*, 2012, pp. 67–72.

[16] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," *18th International Society for Music Information Retrieval Conferenceng, Suzhou, China*, 2017.

[17] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks." in *ISMIR*, 2014, pp. 477–482.

[18] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," *arXiv preprint arXiv:1703.08019*, 2017.

[19] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," *arXiv preprint arXiv:1803.00702*, 2018.

[20] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 61–65.

[21] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 246–250.

[22] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[23] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[24] E. M. Grais, G. Roma, A. J. Simpson, M. D. Plumbley, E. M. Grais, G. Roma, A. J. Simpson, and M. D. Plumbley, "Two-stage single-channel audio source separation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 9, pp. 1773–1783, 2017.

[25] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[27] A. Hyvarinen and H. Morioka, "Unsupervised feature extraction by time-contrastive learning and nonlinear ica," in *Advances in Neural Information Processing Systems*, 2016, pp. 3765–3773.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[29] M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[30] Z. Rafii, A. Liutkus, F.-R. Stter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372