SINGING VOICE SEPARATION WITH DEEP U-NET CONVOLUTIONAL NETWORKS

Andreas Jansson[1, 2], **Eric Humphrey**[2], **Nicola Montecchio**[2],
Rachel Bittner[2], **Aparna Kumar**[2], **Tillman Weyde**[1]
[1]City, University of London, [2]Spotify
{andreas.jansson.1, t.e.weyde}@city.ac.uk
{ejhumphrey, venice, rachelbittner, aparna}@spotify.com

ABSTRACT

The decomposition of a music audio signal into its vocal and backing track components is analogous to image-to-image translation, where a mixed spectrogram is transformed into its constituent sources. We propose a novel application of the *U-Net* architecture — initially developed for medical imaging — for the task of source separation, given its proven capacity for recreating the fine, low-level detail required for high-quality audio reproduction. Through both quantitative evaluation and subjective assessment, experiments demonstrate that the proposed algorithm achieves state-of-the-art performance.

1. INTRODUCTION

The field of Music Information Retrieval (MIR) concerns itself, among other things, with the analysis of music in its many facets, such as melody, timbre or rhythm [20]. Among those aspects, popular western commercial music ("pop" music) is arguably characterized by emphasizing mainly the Melody and Accompaniment aspects; while this is certainly an oversimplification in the context of the whole genre, we restrict the focus of this paper to the analysis of music that lends itself well to be described in terms of a main melodic line (foreground) and accompaniment (background) [27]. Normally the melody is sung, whereas the accompaniment is performed by one or more instrumentalists; a singer delivers the lyrics, and the backing musicians provide harmony as well as genre and style cues [29].

The task of automatic singing voice separation consists of estimating what the sung melody and accompaniment would sound like in isolation. A clean vocal signal is helpful for other related MIR tasks, such as singer identification [18] and lyric transcription [17]. As for commercial applications, it is evident that the karaoke industry, estimated to be worth billions of dollars globally [4], would directly benefit from such technology.

2. RELATED WORK

Several techniques have been proposed for blind source separation of musical audio. Successful results have been achieved with non-negative matrix factorization [26, 30, 32], Bayesian methods [21], and the analysis of repeating structures [23].

Deep learning models have recently emerged as powerful alternatives to traditional methods. Notable examples include [25] where a deep feed-forward network learns to estimate an ideal binary spectrogram mask that represents the spectrogram bins in which the vocal is more prominent than the accompaniment. In [9] the authors employ a deep recurrent architecture to predict soft masks that are multiplied with the original signal to obtain the desired isolated source.

Convolutional encoder-decoder architectures have been explored in the context of singing voice separation in [6] and [8]. In both of these works, spectrograms are compressed through a bottleneck layer and re-expanded to the size of the target spectrogram. While this "hourglass" architecture is undoubtedly successful in discovering global patterns, it is unclear how much local detail is lost during contraction.

One potential weakness shared by the papers cited above is the lack of large training datasets. Existing models are usually trained on hundreds of tracks of lower-than-commercial quality, and may therefore suffer from poor generalization. In this work we aim to mitigate this problem using weakly labeled professionally produced music tracks.

Over the last few years, considerable improvements have occurred in the family of machine learning algorithms known as image-to-image translation [11] — pixel-level classification [2], automatic colorization [33], image segmentation [1] — largely driven by advances in the design of novel neural network architectures.

This paper formulates the voice separation task, whose domain is often considered from a time-frequency perspective, as the translation of a mixed spectrogram into vocal and instrumental spectrograms. By using this framework we aim to make use of some of the advances in image-to-image translation — especially in regard to the reproduc-

tion of fine-grained details — to advance the state-of-the-art of blind source separation for music.

3. METHODOLOGY

This work adapts the *U-Net* [24] architecture to the task of vocal separation. The architecture was introduced in biomedical imaging, to improve precision and localization of microscopic images of neuronal structures. The architecture builds upon the *fully convolutional network* [14] and is similar to the *deconvolutional network* [19]. In a deconvolutional network, a stack of convolutional layers — where each layer halves the size of the image but doubles the number of channels — encodes the image into a small and deep representation. That encoding is then decoded to the original size of the image by a stack of upsampling layers.

In the reproduction of a natural image, displacements by just one pixel are usually not perceived as major distortions. In the frequency domain however, even a minor linear shift in the spectrogram has disastrous effects on perception: this is particularly relevant in music signals, because of the logarithmic perception of frequency; moreover, a shift in the time dimension can become audible as jitter and other artifacts. Therefore, it is crucial that the reproduction preserves a high level of detail. The U-Net adds additional skip connections between layers at the same hierarchical level in the encoder and decoder. This allows low-level information to flow directly from the high-resolution input to the high-resolution output.

3.1 Architecture

The goal of the neural network architecture is to predict the vocal and instrumental components of its input indirectly: the output of the final decoder layer is a soft mask that is multiplied element-wise with the mixed spectrogram to obtain the final estimate. Figure 1 outlines the network architecture. In this work, we choose to train two separate models for the extraction of the instrumental and vocal components of a signal, to allow for more divergent training schemes for the two models in the future.

3.1.1 Training

Let X denote the magnitude of the spectrogram of the original, mixed signal, that is, of the audio containing both vocal and instrumental components. Let Y denote the magnitude of the spectrograms of the target audio; the latter refers to either the vocal (Y_v) or the instrumental (Y_i) component of the input signal.

The loss function used to train the model is the $L_{1,1}$ norm [1] of the difference of the target spectrogram and the masked input spectrogram:

$$L(X, Y; \Theta) = ||f(X, \Theta) \odot X - Y||_{1,1} \qquad (1)$$

where $f(X, \Theta)$ is the output of the network model applied to the input X with parameters Θ – that is the mask generated by the model.

[1] The $L_{1,1}$ norm of a matrix is simply the sum of the absolute values of its elements.

Two U-Nets, Θ_v and Θ_i, are trained to predict vocal and instrumental spectrogram masks, respectively.

3.1.2 Network Architecture Details

Our implementation of U-Net is similar to that of [11]. Each encoder layer consists of a strided 2D convolution of stride 2 and kernel size 5x5, batch normalization, and leaky rectified linear units (ReLU) with leakiness 0.2. In the decoder we use strided deconvolution (sometimes referred to as transposed convolution) with stride 2 and kernel size 5x5, batch normalization, plain ReLU, and use 50% dropout to the first three layers, as in [11]. In the final layer we use a sigmoid activation function. The model is trained using the ADAM [12] optimizer.

Given the heavy computational requirements of training such a model, we first downsample the input audio to 8192 Hz in order to speed up processing. We then compute the Short Time Fourier Transform with a window size of 1024 and hop length of 768 frames, and extract patches of 128 frames (roughly 11 seconds) that we feed as input and targets to the network. The magnitude spectrograms are normalized to the range $[0, 1]$.

3.1.3 Audio Signal Reconstruction

The neural network model operates exclusively on the magnitude of audio spectrograms. The audio signal for an individual (vocal/instrumental) component is rendered by constructing a spectrogram: the output magnitude is given by applying the mask predicted by the U-Net to the magnitude of the original spectrum, while the output phase is that of the original spectrum, unaltered. Experimental results presented below indicate that such a simple methodology proves effective.

3.2 Dataset

As stated above, the description of the model architecture assumes that training data was available in the form of a triplet (original signal, vocal component, instrumental component). Unless one is in the extremely fortunate position as to have access to vast amounts of unmixed multitrack recordings, an alternative strategy has to be found in order to train a model like the one described.

A solution to the issue was found by exploiting a specific but large set of commercially available recordings in order to "construct" training data: *instrumental versions* of recordings.

It is not uncommon for artists to release instrumental versions of tracks along with the original mix. We leverage this fact by retrieving pairs of (original, instrumental) tracks from a large commercial music database. Candidates are found by examining the metadata for tracks with matching duration and artist information, where the track title (fuzzily) matches except for the string "Instrumental" occurring in exactly one title in the pair. The pool of tracks is pruned by excluding exact content matches. Details about the construction of this dataset can be found in [10].

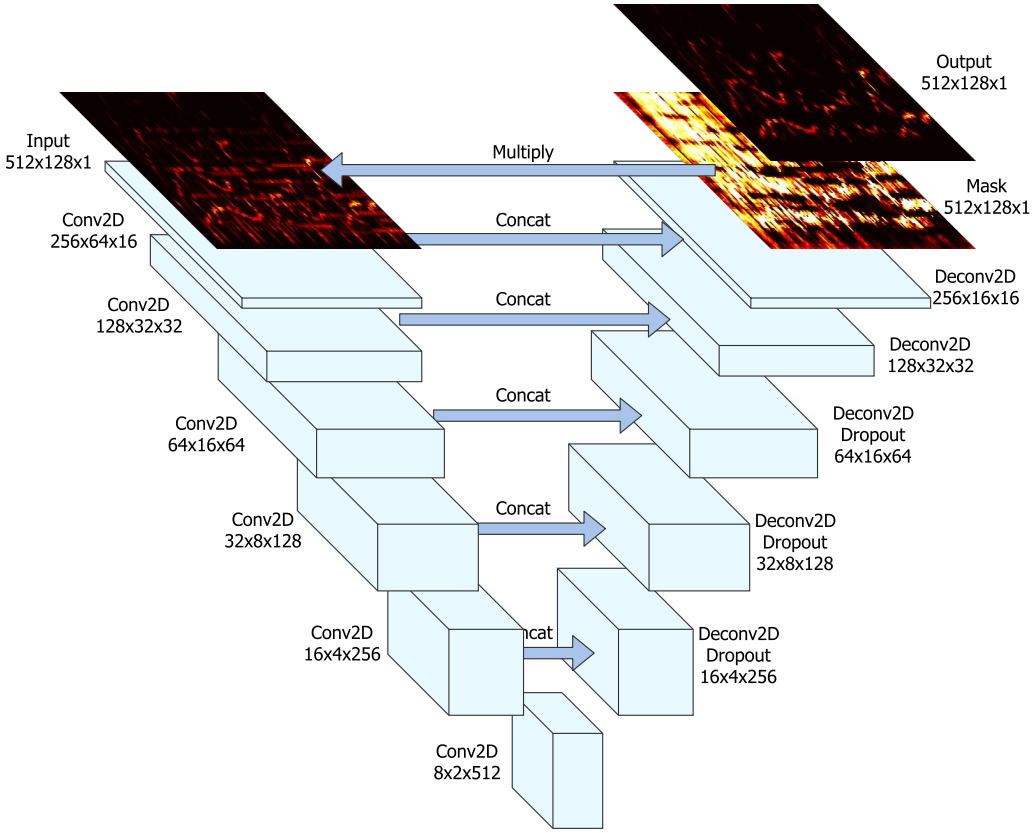

Figure 1. Network Architecture

Genre	Percentage
Pop	26.0%
Rap	21.3%
Dance & House	14.2%
Electronica	7.4%
R&B	3.9%
Rock	3.6%
Alternative	3.1%
Children's	2.5%
Metal	2.5%
Latin	2.3%
Indie Rock	2.2%
Other	10.9%

Table 1. Training data genre distribution

The above approach provides a large source of X (mixed) and Y_i (instrumental) magnitude spectrogram pairs. The vocal magnitude spectrogram Y_v is obtained from their half-wave rectified difference. A qualitative analysis of a large handful of examples showed that this technique produced reasonably isolated vocals.

The final dataset contains approximately 20,000 track pairs, resulting in almost two months worth of continuous audio. To the best of our knowledge, this is the largest training data set ever applied to musical source separation. Table 1 shows the relative distribution of the most frequent genres in the dataset, obtained from the catalog metadata.

4. EVALUATION

We compare the proposed model to the *Chimera* model [15] that produced the highest evaluation scores in the 2016 MIREX Source Separation campaign [2]; we make use of their web interface [3] to process audio clips. It should be noted that the Chimera web server is running an improved version of the algorithm that participated in MIREX, using a hybrid "multiple heads" architecture that combines deep clustering with a conventional neural network [16].

For evaluation purposes we built an additional baseline model; it resembles the U-Net model but without the skip connections, essentially creating a convolutional encoder-decoder, similar to the "Deconvnet" [19].

We evaluate the three models on the standard iKala [5] and MedleyDB dataset [3]. The iKala dataset has been used as a standardized evaluation for the annual MIREX campaign for several years, so there are many existing results that can be used for comparison. MedleyDB on the other hand was recently proposed as a higher-quality, commercial-grade set of multi-track stems. We generate isolated instrumental and vocal tracks by weighting sums of instrumental/vocal stems by their respective mixing co-

[2] www.music-ir.org/mirex/wiki/2016:Singing_Voice_Separation_Results

[3] danetapi.com/chimera

	U-Net	Baseline	Chimera
NSDR Vocal	**11.094**	8.549	8.749
NSDR Instrumental	**14.435**	10.906	11.626
SIR Vocal	**23.960**	20.402	21.301
SIR Instrumental	**21.832**	14.304	20.481
SAR Vocal	**17.715**	15.481	15.642
SAR Instrumental	**14.120**	12.002	11.539

Table 2. iKala mean scores

	U-Net	Baseline	Chimera
NSDR Vocal	**8.681**	7.877	6.793
NSDR Instrumental	**7.945**	6.370	5.477
SIR Vocal	**15.308**	14.336	12.382
SIR Instrumental	**21.975**	16.928	20.880
SAR Vocal	**11.301**	10.632	10.033
SAR Instrumental	**15.462**	15.332	12.530

Table 3. MedleyDB mean scores

Model	Mean	SD	Min	Max	Median
U-Net	**14.435**	3.583	4.165	21.716	**14.525**
Baseline	10.906	3.247	1.846	19.641	10.869
Chimera	11.626	4.151	-0.368	20.812	12.045
LCP2	11.188	3.626	2.508	19.875	11.000
LCP1	10.926	3.835	0.742	19.960	10.800
MC2	9.668	3.676	-7.875	22.734	9.900

Table 4. iKala NSDR Instrumental, MIREX 2016

Model	Mean	SD	Min	Max	Median
U-Net	**11.094**	3.566	2.392	20.720	**10.804**
Baseline	8.549	3.428	-0.696	18.530	8.746
Chimera	8.749	4.001	-1.850	18.701	8.868
LCP2	6.341	3.370	-1.958	17.240	5.997
LCP1	6.073	3.462	-1.658	17.170	5.649
MC2	5.289	2.914	-1.302	12.571	4.945

Table 5. iKala NSDR Vocal, MIREX 2016

efficients as supplied by the MedleyDB Python API [4]. We limit our evaluation to clips that are known to contain vocals, using the melody transcriptions provided in both iKala and MedleyDB.

The following functions are used to measure performance: Signal-To-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR) [31]. Normalized SDR (NSDR) is defined as

$$\text{NSDR}(S_e, S_r, S_m) = \text{SDR}(S_e, S_r) - \text{SDR}(S_m, S_r) \quad (2)$$

where S_e is the estimated isolated signal, S_r is the reference isolated signal, and S_m is the mixed signal. We compute performance measures using the *mir_eval* toolkit [22].

Table 2 and Table 3 show that the U-Net significantly outperforms both the baseline model and Chimera on all three performance measures for both datasets. In Figure 2 we show an overview of the distributions for the different evaluation measures.

Assuming that the distribution of tracks in the iKala hold-out set used for MIREX evaluations matches those in the public iKala set, we can compare our results to the participants in the 2016 MIREX Singing Voice Separation task. [5] Table 4 and Table 5 show NSDR scores for our models compared to the best performing algorithms of the 2016 MIREX campaign.

In order to assess the effect of the U-Net's skip connections, we can visualize the masks generated by the U-Net and baseline models. From Figure 3 it is clear that while the baseline model captures the overall structure, there is a lack of fine-grained detail observable.

4.1 Subjective Evaluation

Emiya et al. introduced a protocol for the subjective evaluation of source separation algorithms [7]. They suggest asking human subjects four questions that broadly correspond to the SDR/SIR/SAR measures, plus an additional question regarding the overall sound quality.

As we asked these four questions to subjects without music training, our subjects found them ambiguous, e.g., they had problems discerning between the absence of artifacts and general sound quality. For better clarity, we distilled the survey into the following two questions in the vocal extraction case:

- Quality: "Rate the vocal quality in the examples below."
- Interference: "How well have the instruments in the clip above been removed in the examples below?"

For instrumental extraction we asked similar questions:

- Quality: "Rate the sound quality of the examples below relative to the reference above."
- Extracting instruments: "Rate how well the instruments are isolated in the examples below relative to the full mix above."

Data was collected using CrowdFlower [6], an online platform where humans carry out micro-tasks, such as image classification, simple web searches, etc., in return for small per-task payments.

In our survey, CrowdFlower users were asked to listen to three clips of isolated audio, generated by U-Net, the baseline model, and Chimera. The order of the three clips was randomized. Each question asked one of the Quality and Interference questions. In the Interference question we also included a reference clip. The answers were given according to a 7 step Likert scale [13], ranging from "Poor" to "Perfect". Figure 4 is a screen capture of a CrowdFlower question.

[4] github.com/marl/medleyDB

[5] http://www.music-ir.org/mirex/wiki/2016:Singing_Voice_Separation_Results

[6] www.crowdflower.com

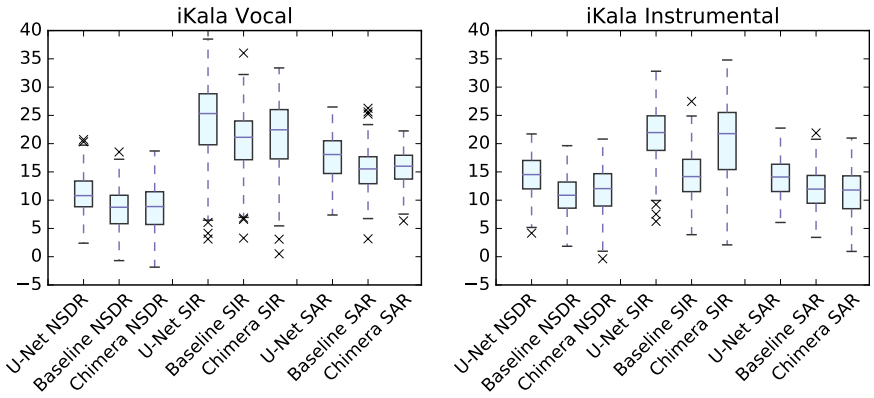

Figure 2. iKala vocal and instrumental scores

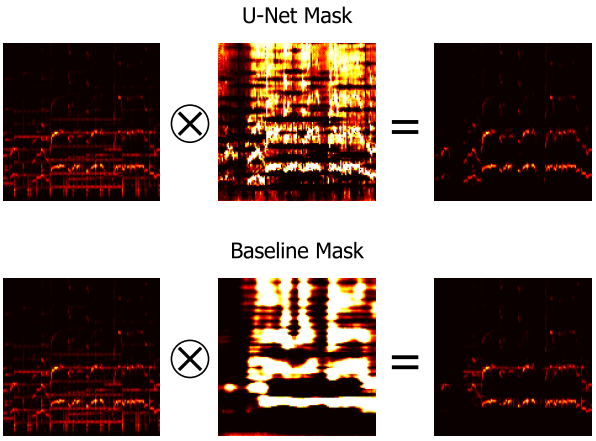

Figure 3. U-Net and baseline masks

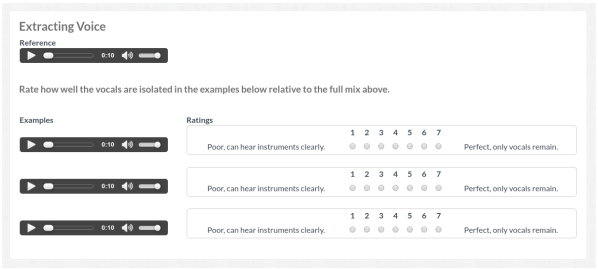

Figure 4. CrowdFlower example question

To ensure the quality of the collected responses, we interspersed the survey with "control questions" that the user had to answer correctly according to a predefined set of acceptable answers on the Likert scale. Users of the platform are unaware of which questions are control questions. If they are answered incorrectly, the user is disqualified from the task. A music expert external to our research group was asked to provide acceptable answers to a number of random clips that were designated as control questions.

For the survey we used 25 clips from the iKala dataset and 42 clips from MedleyDB. We had 44 respondents and 724 total responses for the instrumental test, and 55 respondents supplied 779 responses for the voice test [7].

Figure 5 shows mean and standard deviation for answers provided on CrowdFlower. The U-Net algorithm outperforms the other two models on all questions.

5. CONCLUSION AND FUTURE WORK

We have explored the U-Net architecture in the context of singing voice separation, and found that it brings clear improvements over the state-of-the-art. The benefits of low-level skip connections were demonstrated by comparison to plain convolutional encoder-decoders.

A factor that we feel should be investigated further is the impact of large training data: work remains to be done to correlate the effects of the size of the training dataset to the quality of source separation.

We have observed some examples of poor separation on tracks where the vocals are mixed at lower-than-average volume, uncompressed, suffer from extreme application of audio effects, or otherwise unconventionally mixed. Since the training data consisted exclusively of commercially produced recordings, we hypothesize that our model has learned to distinguish the kind of voice typically found in commercial pop music. We plan to investigate this further by systematically analyzing the dependence of model performance on the mixing conditions.

Finally, subjective evaluation of source separation algorithms is an open research question. Several alternatives exist to 7-step Likert scale, e.g. the ITU-R scale [28]. Tools like CrowdFlower allow us to quickly roll out surveys, but care is required in the design of question statements.

6. REFERENCES

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[7] Some of the audio clips we used for evaluation can be found on `http://mirg.city.ac.uk/codeapps/vocal-source-separation-ismir2017`

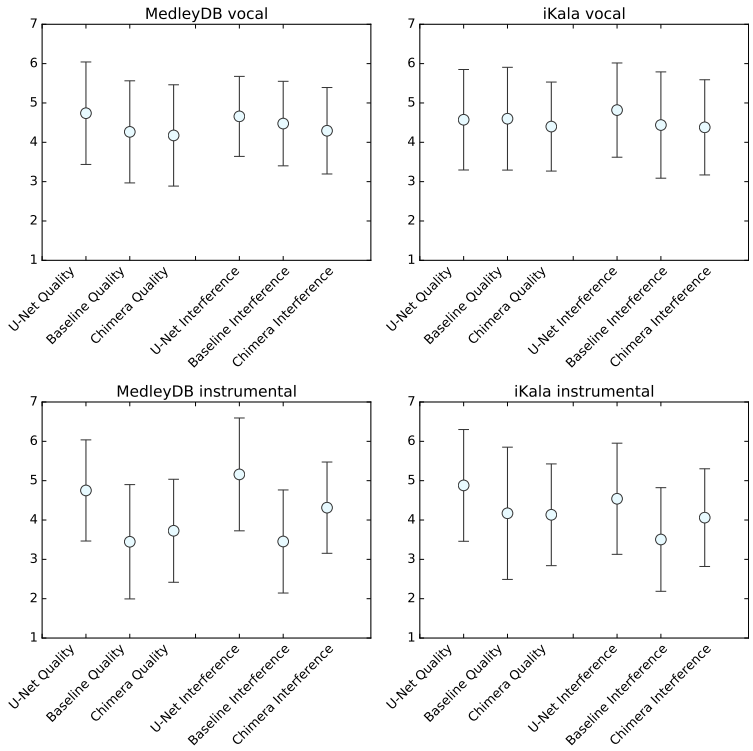

Figure 5. CrowdFlower evaluation results (mean/std)

[2] Aayush Bansal, Xinlei Chen, Bryan Russell, Abhinav Gupta, and Deva Ramanan. Pixelnet: Towards a general pixel-level architecture. *arXiv preprint arXiv:1609.06694*, 2016.

[3] Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, pages 155–160, 2014.

[4] Kevin Brown. *Karaoke Idols: Popular Music and the Performance of Identity*. Intellect Books, 2015.

[5] Tak-Shing Chan, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang, and Roger Jang. Vocal activity informed singing voice separation with the iKala dataset. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 718–722. IEEE, 2015.

[6] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 258–266. Springer, 2017.

[7] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2046–2057, 2011.

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

[9] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Singing-voice separation from monaural recordings using deep recurrent neural networks. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, pages 477–482, 2014.

[10] Eric Humphrey, Nicola Montecchio, Rachel Bittner, Andreas Jansson, and Tristan Jehan. Mining labeled data from web-scale collections for vocal activity detection in music. In *Proceedings of the 18th ISMIR Conference*, 2017.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[12] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[15] Yi Luo, Zhuo Chen, and Daniel PW Ellis. Deep clustering for singing voice separation. 2016.

[16] Yi Luo, Zhuo Chen, John R Hershey, Jonathan Le Roux, and Nima Mesgarani. Deep clustering and conventional networks for music separation: Stronger together. *arXiv preprint arXiv:1611.06265*, 2016.

[17] Annamaria Mesaros and Tuomas Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(1):546047, 2010.

[18] Annamaria Mesaros, Tuomas Virtanen, and Anssi Klapuri. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23-27, 2007*, pages 375–378, 2007.

[19] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[20] Nicola Orio et al. Music retrieval: A tutorial and review. *Foundations and Trends® in Information Retrieval*, 1(1):1–90, 2006.

[21] Alexey Ozerov, Pierrick Philippe, Frdric Bimbot, and Rmi 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*, 15(5):1564–1578, 2007.

[22] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. Mir_eval: A transparent implementation of common MIR metrics. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, pages 367–372, 2014.

[23] Zafar Rafii and Bryan Pardo. Repeating pattern extraction technique (REPET): A simple method for music/voice separation. *IEEE transactions on audio, speech, and language processing*, 21(1):73–84, 2013.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[25] Andrew JR Simpson, Gerard Roma, and Mark D Plumbley. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 429–436. Springer, 2015.

[26] Paris Smaragdis, Cedric Fevotte, Gautham J Mysore, Nasser Mohammadiha, and Matthew Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, 2014.

[27] Philip Tagg. Analysing popular music: theory, method and practice. *Popular music*, 2:37–67, 1982.

[28] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and Catherine Colomes. Peaq-the itu standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–29, 2000.

[29] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

[30] Shankar Vembu and Stephan Baumann. Separation of vocals from polyphonic audio recordings. In *ISMIR 2005, 6th International Conference on Music Information Retrieval, London, UK, 11-15 September 2005, Proceedings*, pages 337–344, 2005.

[31] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

[32] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007.

[33] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.