Interactive Deep Singing-Voice Separation Based on Human-in-the-Loop Adaptation

Tomoyasu Nakano
t.nakano@aist.go.jp
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

Yuki Koyama
koyama.y@aist.go.jp
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

Masahiro Hamasaki
masahiro.hamasaki@aist.go.jp
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

Masataka Goto
m.goto@aist.go.jp
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

ABSTRACT

This paper presents a deep-learning-based interactive system separating the singing voice from input polyphonic music signals. Although deep neural networks have been successful for singing voice separation, no approach using them allows any user interaction for improving the separation quality. We present a framework that allows a user to interactively fine-tune the deep neural model at run time to adapt it to the target song. This is enabled by designing unified networks consisting of two U-Net architectures based on frequency spectrogram representations: one for estimating the spectrogram mask that can be used to extract the singing-voice spectrogram from the input polyphonic spectrogram; the other for estimating the fundamental frequency (F0) of the singing voice. Although it is not easy for the user to edit the mask, he or she can iteratively correct errors in part of the visualized F0 trajectory through simple interaction. Our unified networks leverage the user-corrected F0 to improve the rest of the F0 trajectory through the model adaptation, which results in better separation quality. We validated this approach in a simulation experiment showing that the F0 correction can improve the quality of singing-voice separation. We also conducted a pilot user study with an expert musician, who used our system to produce a high-quality singing-voice separation result.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**.

KEYWORDS

Singing-voice separation, deep learning, human-in-the-loop model adaptation, F0 estimation

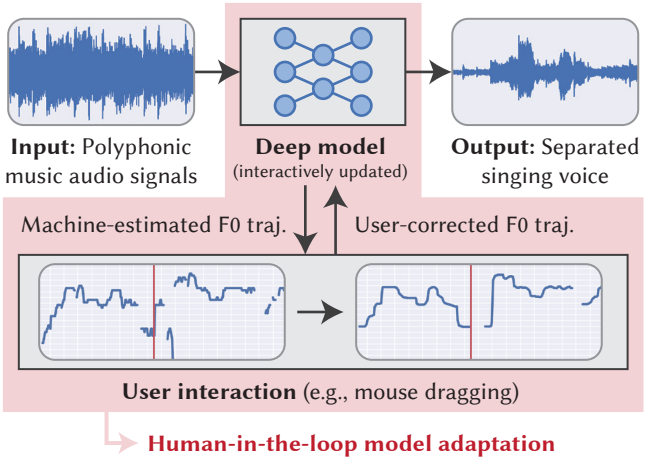

Figure 1: Overview of *interactive deep singing-voice separation*, where the user can interact with the deep neural network model to fine-tune the quality of the separated singing voice. The user feedback is provided as F0 trajectories, based on which the model is adapted.

ACM Reference Format:
Tomoyasu Nakano, Yuki Koyama, Masahiro Hamasaki, and Masataka Goto. 2020. Interactive Deep Singing-Voice Separation Based on Human-in-the-Loop Adaptation. In *25th International Conference on Intelligent User Interfaces (IUI '20), March 17–20, 2020, Cagliari, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3377325.3377539

1 INTRODUCTION

As the use of subscription music services has become commonplace, efficient music information retrieval systems and intelligent music listening interfaces have become more and more important to increase the value of human music activities. Since singing voice is one of the most important elements in music [17] and many listeners pay attention to the singing voice and its lyrics [6], *singing-voice separation* from polyphonic music signals [30] has been a fundamental technology for various real-world applications. It can, for example, be used to retrieve songs with singers having voice timbre similar to that of a listener's favorite artist [13], to improve the performance of audio-to-lyrics synchronization for automatically

IUI '20, March 17–20, 2020, Cagliari, Italy

generating lyric videos [20], and to provide an intelligent music-playback interface that allows the user to interactively change the playback order of multiple songs according to vocal gender [26].

The majority of recently proposed singing-voice separation methods involve deep learning [11, 16, 19, 23, 24, 30, 32–34], where researchers have mainly focused on how to improve the accuracy of singing-voice separation as much as possible for songs as various as possible. Thus the main concern is to seek better training methods, better model architectures, and larger training datasets that achieve better accuracy with better generalization ability. It is, however, difficult for even state-of-the-art methods to achieve the perfect separation (*e.g.*, electric guitar sounds in some songs are sometimes recognized as singing voice), and there is no means to improve the quality except for feeding more training data, which is not a feasible option for end-users. Thus, with these methods it is not easy to obtain a high-quality separation result for a specific song even if a high interaction cost can be afforded.

In this work, we take a different approach: instead of using a static *general* model for separation, we investigate fine-tuning of an initial model based on U-Net architectures to adapt it to a *specific* target song to improve the separation accuracy for that song. Specifically, we propose a human-in-the-loop model adaptation framework (see Figure 1) where the user iteratively provides feedback to the model for fine-tuning for the target song. We call this framework *interactive deep singing-voice separation*. The key idea is to perform this feedback loop in the domain of *fundamental frequency* (F0, which is a physical property of sound related to pitch) of the singing voice, instead of the domain of separated audio signals or their frequency spectrograms. This allows the user to easily communicate with the model through simple interaction (*e.g.*, mouse dragging) for correcting errors in the visualized F0 trajectory. In this framework, the user does not have to correct all the F0 errors. After correcting some of them, the user can use them to update the deep source separation model by adapting it to the target song, which could be expected to result in automatically correcting some remaining errors. This could reduce the burden of manually correcting errors to achieve the perfect separation.

Note that our approach is complementary to the approach of training a general model; our framework could use any state-of-the-art training methods, architectures, and datasets in the future to train an initial model, and then let users interactively adapt the model for a specific song to obtain even better separation.

Our contributions are summarized as follows:

- We present the first singing-voice separation framework that allows users to interactively fine-tune a deep-learning model to a target song to increase separation accuracy.
- We propose to let users interact with estimated F0 trajectories instead of separated singing voices to provide feedback to the model since it is much easier for humans to edit F0 trajectory than to edit audio signals.
- We validated through simulated experiments that our framework could improve the separation accuracy by feeding corrected F0 trajectories. We also validated through a pilot user study that an expert musician could perform interactive model adaptation with our system and then obtain a better singing-voice separation result.

2 RELATED WORK

User-guided sound source separation methods have already been investigated, though they are not based on deep learning and do not focus on singing voices. Those methods can leverage human interaction to improve the accuracy of sound source separation. For example, by giving an audio example such as a singing or speaking voice as side information, the corresponding sound source that is similar to the example can be separated [25, 31]. The separation accuracy can be improved by annotating the time when each sound source exists alone [28], annotating the F0 by hand [9, 12], and annotating sound sources in the time-frequency domain [4, 7, 22]. These methods, however, have not shown how to use human interaction to improve the separation accuracy of deep learning. In contrast, we investigate how users can be involved in the deep-learning-based separation process, and we show how to adapt the deep model by using the F0 error correction.

Several methods that allow users to interactively fine-tune deep neural models have been proposed in domains other than singing-voice separation. For example, in the image processing domain, researchers have proposed interactive image segmentation methods that allow interactive fine-tuning [35, 36]. In this work, we present the first interactive deep method for the singing-voice separation problem and propose the novel idea of using the F0 trajectory as the interface between the deep neural model and the user.

Our framework is related to the approach called *interactive machine learning* [1, 10] in that training of machine learning models happens during user interaction. While the main purpose of interactive machine learning is to allow an end-user to rapidly train a general model to the target problem, our framework uses interactivity for training a specific model that works well to a specific instance of the target problem.

3 INTERACTIVE DEEP SINGING-VOICE SEPARATION

We propose a deep-learning-based singing-voice separation method that utilizes user guides. Our key idea is to use the fundamental frequency (F0) of singing voice, which is an acoustic feature of singing voice and is closely related to the improvement of singing-voice separation quality, as the interface for the user to interact with the deep model since it is much easier for users to correct F0 trajectories than to correct separated singing voice signals. Performance improvements of F0 estimation and singing-voice separation in music have been used to improve each other's performance [5, 8, 15]. Recently, deep-learning-based methods that jointly estimate singing-voice F0 and separated singing voice have also been proposed [18, 27].

Figure 2 shows our prototype system, which displays the input polyphonic music audio signals, the estimated or user-corrected F0 trajectory, and the separated singing voice. The F0 estimation error can be easily corrected by simple mouse-based interactions and is used as a user's guide to update separated singing voice. The user corrects F0 by dragging the mouse, uses the left button to draw new F0 trajectory, and uses the right button to delete existing F0 trajectory. The parameters of the deep-learning model are tuned by pressing the "Update DNN" button after correcting the F0 trajectory. Since our current implementation assumes that the user corrects

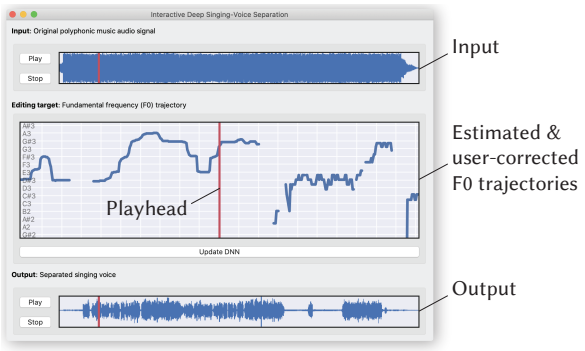

Figure 2: Screen capture of our prototype system. (Top) Input polyphonic music audio signals. (Middle) Widget for correcting the estimated F0 trajectory. (Bottom) Current output of the separated-singing voice.

F0 from the beginning, the parameters are updated with the data up to the point when the user last corrected it.

By tuning the model parameters, two effects can be obtained. The first is to speed up the manual process of the F0 error correction. Since the F0 of the uncorrected part can be re-estimated by the deep learning model that is updated by using the partially corrected F0, F0 error correction can be speeded up without correcting all F0 trajectories. Although such a function does not exist in the previous work on interactive time-pitch annotation-informed separation [12], this kind of interaction was proposed for acoustic event detection [21]. The second effect obtained by tuning the model parameters is to improve the singing-voice separation quality. The user can obtain a singing voice with improved separation quality by simply deleting the F0 where the singing voice does not exist and, elsewhere, by drawing the correct F0 of the singing voice part.

4 IMPLEMENTATION

4.1 Deep Singing-Voice Separation

We use the U-Net architectures for both the singing-voice separation and the F0 estimation. Based on Jansson *et al.* [19], amplitude spectra obtained using the short-time Fourier transform (STFT) are used for the input to the U-Net. They are computed with the LibROSA python library from a monaural audio signal (with a 44.1 kHz sampling rate) that is obtained by averaging left and right channels in the original stereo audio signal of a song. The STFT conditions and the U-Net architecture are the same as in [27] except that the U-Net has only one decoder that outputs a vocal mask.

The output of the U-Net for the F0 estimation is represented as a two-dimensional F0 saliency map (corresponding to the frequency spectrogram) [2], in which the F0 saliency value is between 0 and 1 at each time-frequency bin. The ground truth of the F0 saliency map is obtained as a Gaussian-blurred binary saliency spectrogram (*i.e.*, quantized F0 trajectory) as proposed in [2].

In order to have the user's F0 correction result contribute to the improvement of the singing voice separation quality, we connect the U-Net for the source separation and the U-Net for the F0 estimation and estimate their parameters jointly. There are various ways of connecting the U-Nets. Jansson *et al.* [18] reported that a joint

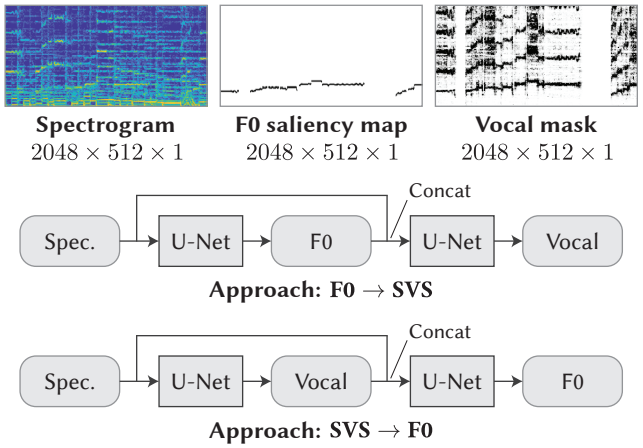

Figure 3: Overview of our model. Two U-Nets are concatenated to estimate both a vocal mask (SVS) and an F0 saliency map (F0). We have two variants; the first one separates singing voice after F0 estimation (F0 → SVS) and the second one separates singing voice first (SVS → F0).

architecture repeating twice a network comprising the U-Net for the source separation followed by the U-Net for the F0 estimation (*i.e.*, four U-Nets in total after the repetition) achieved the best performance. In order to reduce the computation time to improve the user interactivity for our purpose, we implemented two types of non-repetitive connection (Figure 3), one of which concatenates the F0 estimation result to the U-Net input for the singing voice separation (*i.e.*, the U-Net for the F0 estimation followed by the U-Net for the source separation) and the other of which concatenates the singing-voice separation result to the U-Net input for the F0 estimation (*i.e.*, the U-Net for the source separation followed by the U-Net for the F0 estimation).

The whole connected network consisting of the two U-Nets can be trained jointly by the back-propagation algorithm. Since the user-corrected F0 can be used to compute the loss for the back-propagation algorithm, the network parameters (*i.e.*, the deep model) can be updated to adapt to the target song and its singing voice. The connected network of the former type (F0 estimation followed by singing-voice separation) directly updates the U-Net parameters for the singing-voice separation by reflecting the user-corrected F0 and also re-estimates the unmodified F0. In contrast, the connected network of the latter type (singing-voice separation followed by F0 estimation) allows all the parameters of both of the U-Nets to be updated by using the user-corrected F0.

4.2 Interaction

The F0 saliency map is first converted to the F0 trajectory because it is hard for the user to directly correct the map. This conversion is implemented by picking, at each time-frame, the maximum peak value in the F0 saliency map. If the peak value does not exceed the threshold (0.1), the corresponding point on the map is considered mute (silence). This method does not consider the case where there are multiple simultaneous singing voices (multiple F0s at the same time) since this paper focuses on a simple interaction to realize the proposed concept.

As shown in Figure 2, the beat timings and the frequencies corresponding to the note names (*e.g.*, C3 and A3) are visualized as grid lines in order to support correction by the user. Beat timing was automatically estimated using the madmom python library.

5 EVALUATION

We conducted a simulation experiment assuming that the user gave an ideal F0, and we conducted a pilot study with an expert musician.

To evaluate the singing-voice separation accuracy, the MedleyDB dataset [3] and the RWC Music Database [14] were used. The data consisted of 150 songs, 50 from the former and 100 from the latter. From the MedleyDB dataset, we chose 50 songs that contain singing voices with a melody role and have F0 annotations. From neither the MedleyDB dataset nor the RWC Music Database did we choose songs in which the vocal track contained other sounds such as accompaniment piano sounds and the other tracks partially contained some singing voices due to the recording environment.

5.1 Simulation Experiment

To evaluate the separation accuracy by 5-fold cross validation, the 150 songs were randomly divided into five 30-song groups. The standard evaluation tool *mir_eval* [29] was used for computing the signal-to-distortion ratio (SDR) for the singing-voice separation. The SDR was calculated for each song, and then its median over all songs was calculated.

By using the training set of the cross validation, the model parameters were first trained for 256 epochs (iterations). Then, to simulate a situation where a user partially or fully corrected F0 trajectories, we partially or fully used the correct F0 annotation of each song in the training set to fine-tune the parameters for 32 epochs. From the beginning of each song, we used 25%, 50%, 75%, or 100% of the annotation.

We compared the performances of the following conditions:

- **SVS (0%):** A single U-Net for the singing-voice separation (without using the F0 estimation).
- **SVS → F0 (0%):** The U-Net for the singing-voice separation followed by the U-Net for the F0 estimation[1].
- **F0 (0%) → SVS:** The U-Net for the F0 estimation followed by the U-Net for the singing-voice separation.
- **SVS → F0 (25–100%)** or **F0 (25–100%) → SVS:** Our deep singing-voice separation with the model adaptation.

Note that the percentage value represents the ratio of the correct F0 annotation used for the adaptation (fine-tuning). 0% means that none of the F0 annotation is used and 100% means that all the F0 annotation of each song is used.

Figure 4 shows the results. The network that separates singing voice after F0 estimation (F0 → SVS) had lower separation quality than the network that separates singing voice first (SVS → F0) when no model adaptation was performed (0%), but it had higher separation quality when a sufficient amount (50%, 75%, and 100%) of the correct F0 annotation (*i.e.*, the user's F0 correction) was given. In addition, even when the F0 correction was performed only partially (F0 (50%) → SVS), the result showed higher performance than the case without the F0 correction. Furthermore, our method with the sufficient model adaption, F0 (100%) → SVS, greatly outperformed the methods without the model adaptation (0%).

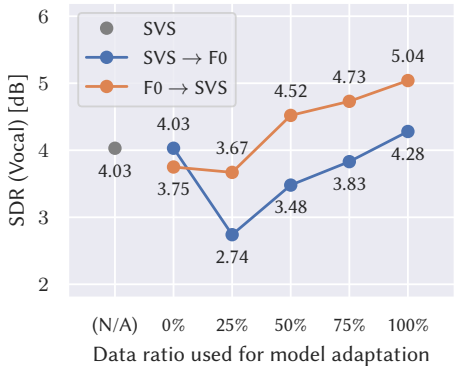

Figure 4: SDR value (higher is better) in each condition.

5.2 Pilot User Study

We also conducted an informal pilot study to see whether our main target users (*i.e.*, expert musicians who are familiar with the concept of F0) could correct estimated F0 trajectories and fine-tune the model to obtain better results. The participant was an expert musician who had experience in correcting estimated F0 trajectories of singing voice using existing software, Audionamix TRAX PRO 3. We asked the participant to separate one song, RWC-MDB-P No. 007 [14]. We asked the participant to first check the initial F0 estimation and singing-voice separation results and then begin the iteration of F0 correction and model adaptation for 30 minutes.

In an informal interview after this session, the participant commented that the fine-tuning to automatically improve the rest of the area by model adaptation was effective. For example, we observed that errors such as a guitar solo mistakenly separated as singing voice were automatically corrected by model adaptation with F0 correction of other parts.

6 CONCLUSION

This paper presented an interactive deep-learning-based framework that can separate the singing voice from input polyphonic music signals. The key idea was to involve the user in the loop to obtain feedback about the target song, based on which the deep model can be fine-tuned. The simulation experiment showed that our framework could improve singing-voice separation quality. In addition, we have shown in a user study that a user successfully corrected F0 trajectories to improve singing-voice separation quality.

The current interaction for F0 correction is a simple mouse operation, and a future challenge is to improve usability with intelligent support (*e.g.*, like that provided by a magnetic selection tool used in image-editing software and some commercial singing-voice separation software). We also plan to investigate the use of information other than F0 for even better human-in-the-loop singing-voice separation.

ACKNOWLEDGMENTS

This work was supported in part by JST ACCEL (JPMJAC1602).

[1] Note that our connection method is different from that used in previous research [18].

REFERENCES

[1] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (Dec. 2014), 105–120.

[2] R. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello. 2017. Deep Salience Representations for F0 Estimation in Polyphonic Music. In *ISMIR 2017*. 63–70.

[3] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello. 2014. MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In *ISMIR 2014*. 155–160.

[4] J. Bryan, Gautham J. Mysore, and Ge Wang. 2014. ISSE: An Interactive Source Separation Editor. In *ACM CHI 2014*.

[5] P. Cabañas-Molero, D. Martínez Muñoz, M. Cobos, and J. J. López. 2011. Singing Voice Separation from Stereo Recordings using Spatial Clues and Robust F0 Estimation. In *AEC Conference*.

[6] A. Demetriou, A. Jansson, A. Kumar, and R. M. Bittner. 2018. Vocals in Music Matter: the Relevance of Vocals in the Minds of Listeners. In *ISMIR 2018*. 514–520.

[7] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot. 2014. An Interactive Audio Source Separation Framework based on Non-negative Matrix Factorization. In *ICASSP 2014*.

[8] J. Durrieu, B. David, and G. Richard. 2011. A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation. *IEEE J. Sel. Topics Signal Process.* 5, 6 (2011), 1180–1191.

[9] J.-L. Durrieu and J.-P. Thiran. 2012. Musical Audio Source Separation based on User-selected F0 Track. In *LVA/ICA 2012*. 438–445.

[10] J. A. Fails and Jr. D. R. Olsen. 2003. Interactive Machine Learning. In *ACM IUI 2003*.

[11] Z. C. Fan, J. S. R. Jang, and C. L. Lu. 2016. Singing Voice Separation and Pitch Extraction from Monaural Polyphonic Audio Music via DNN and Adaptive Pitch Tracking. In *BigMM 2016*. 178–185.

[12] B. Fuentes, R. Badeau, and G. Richard. 2012. Blind Harmonic Adaptive Decomposition Applied to Supervised Source Separation. In *EUSIPCO 2012*.

[13] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno. 2010. A Modeling of Singing Voice Robust to Accompaniment Sounds and its Application to Singer Identification and Vocal-timbre-similarity-based Music Information Retrieval. *IEEE Trans. on Audio, Speech, and Language Processing* 18, 3 (2010), 638–648.

[14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. 2002. RWC Music Database: Popular, Classical, and Jazz Music Databases. In *ISMIR 2002*. 287–288.

[15] C. L. Hsu, D. Wang, J. R. Jang, and K. Hu. 2012. A Tandem Algorithm for Singing Pitch Extraction and Voice Separation From Music Accompaniment. *IEEE Trans. Acoust., Speech, Signal Process.* 20, 5 (2012), 1482–1491.

[16] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. 2014. Singing-voice Separation from Monaural Recordings using Deep Recurrent Neural Networks. In *ISMIR 2014*. 477–482.

[17] E. J. Humphrey, S. Reddy, P. Seetharaman, A. Kumar, R. M. Bittner, A. Demetriou, S. Gulati, A. Jansson, T. Jehan, B. Lehner, A. Krupse, and L. Yang. 2019. An Introduction to Signal Processing for Singing-Voice Analysis: High Notes in the Effort to Automate the Understanding of Vocals in Music. *IEEE Signal Processing Magazine* 36, 1 (2019), 82–94.

[18] A. Jansson, R. M. Bittner, S. Ewert, and T. Weyde. 2019. Joint Singing Voice Separation and F0 Estimation with Deep U-Net Architectures. In *EUSIPCO 2019*.

[19] A. Jansson, E. J. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde. 2017. Singing Voice Separation with Deep U-Net Convolutional Networks. In *ISMIR 2017*. 745–751.

[20] J. Kato, T. Nakano, and M. Goto. 2015. TextAlive: Integrated Design Environment for Kinetic Typography. In *ACM CHI 2015*.

[21] B. Kim and B. Pardo. 2018. A Human-in-the-Loop System for Sound Event Detection and Annotation. *TiiS* 8, 2 (2018), 13:1–13:23.

[22] A. Lefevre, F. Bach, and C. Fevotte. 2012. Semi-supervised NMF with Time-frequency Annotations for Single-channel Source Separation. In *ISMIR 2012*.

[23] K. W. E. Lin, B. T. Balamurali, E. K., S. L., and D. Herremans. 2018. Singing Voice Separation using a Deep Convolutional Neural Network Trained by Ideal Binary Mask and Cross Entropy. *Neural Computing and Applications* (2018), 1–14.

[24] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani. 2017. Deep Clustering and Conventional Networks for Music Separation: Stronger Together. In *ICASSP 2017*. 61–65.

[25] L. Le Magoarou, A. Ozerov, and N. Duong. 2013. Text-Informed Audio Source Separation using Nonnegative Matrix Partial Co-Factorization. In *MLSP 2013*.

[26] T. Nakano, J. Kato, M. Hamasaki, and M. Goto. 2016. PlaylistPlayer: An Interface Using Multiple Criteria to Change the Playback Order of a Music Playlist. In *ACM IUI 2016*. 186–190.

[27] T. Nakano, K. Yoshii, Y. Wu, R. Nishikimi, K. W. E. Lin, and M. Goto. 2019. Joint Singing Pitch Estimation and Voice Separation Based on a Neural Harmonic Structure Renderer. In *IEEE WASPAA 2019*.

[28] A. Ozerov, C. Fevotte, R. Blouet, and J.-L. Durrieu. 2011. Multichannel Nonnegative Tensor Factorization with Structured Constraints for User-guided Audio Source Separation. In *ICASSP 2011*.

[29] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, , and D. P. W. Ellis. 2014. mir_eval: A Transparent Implementation of Common MIR Metrics. In *ISMIR 2014*. 367–372.

[30] Z. Rafii, A. Liutkus, F-R Stoter, S.I. Mimilakis, D. FitzGerald, and B. Pardo. 2018. An Overview of Lead and Accompaniment Separation in Music. *IEEE/ACM Transactions on Audio, Speech, Language Processing* 26, 8 (2018), 1307–1335.

[31] P. Smaragdis and G. Mysore. 2009. "Separation by Humming": User Guided Sound Extraction from Monophonic Mixtures. In *WASPAA 2009*.

[32] D. Stoller, S. Ewert, and S. Dixon. 2017. Wave-U-Net: A Multi-scale Neural Network for End-to-end Audio Source Separation. In *ISMIR 2017*. 330–340.

[33] D. Stoller, S. Ewert, and S. Dixon. 2018. Jointly Detecting and Separating Singing Voice: A Multi-Task Approach. In *LVA/ICA 2018*. 329–339.

[34] N. Takahashi and Y. Mitsufuji. 2017. Multi-scale Multi-band DenseNets for Audio Source Separation. In *WASPAA 2017*. 21–25.

[35] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren. 2018. Interactive Medical Image Segmentation Using Deep Learning With Image-Specific Fine Tuning. *IEEE Trans. on Medical Imaging* 32, 7 (2018), 1562–1573.

[36] N. Xu, B. L. Price, S. Cohen, J. Yang, and T. S. Huang. 2016. Deep Interactive Object Selection. In *CVPR 2016*.