JOINT SINGING PITCH ESTIMATION AND VOICE SEPARATION BASED ON A NEURAL HARMONIC STRUCTURE RENDERER

Tomoyasu Nakano[1] *Kazuyoshi Yoshii*[2] *Yiming Wu*[2] *Ryo Nishikimi*[2]
Kin Wah Edward Lin[1] *Masataka Goto*[1]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Japan
{t.nakano, edward.lin, m.goto}@aist.go.jp
[2]Kyoto University, Japan
{yoshii, wu, nishikimi}@sap.ist.i.kyoto-u.ac.jp

ABSTRACT

This paper describes a multi-task learning approach to joint extraction (fundamental frequency (F0) estimation) and separation of singing voices from music signals. While deep neural networks have been used successfully for each task, both tasks have not been dealt with simultaneously in the context of deep learning. Since vocal extraction and separation are considered to have a mutually beneficial relationship, we propose a unified network that consists of a deep convolutional neural network for vocal F0 saliency estimation and a U-Net with an encoder shared by two decoders specialized for separating vocal and accompaniment parts, respectively. Between these two networks we introduce a differentiable layer that converts an F0 saliency spectrogram into *harmonic* masks indicating the locations of harmonic partials of a singing voice. The physical meaning of harmonic structure is thus reflected in the network architecture. The harmonic masks are then effectively used as scaffolds for estimating *fine-structured* masks thanks to the excellent capability of the U-Net for domain-preserving conversion (*e.g.*, image-to-image conversion). The whole network can be trained jointly by backpropagation. Experimental results showed that the proposed unified network outperformed the conventional independent networks for vocal extraction and separation.

Index Terms— Melody extraction, F0 estimation, singing voice separation, deep learning, multi-task learning

1. INTRODUCTION

A singing voice is one of the most influential elements of music [1]. Accordingly, the estimation of its fundamental frequency (F0) (*a.k.a.* vocal extraction or melody extraction) [2] and singing voice separation (*a.k.a.* vocal separation) [3] have been actively investigated in the field of music information retrieval (MIR). The state-of-the-art studies have successfully used deep neural networks (DNNs) for vocal extraction [4–7] and separation [8–16]. Bittner *et al.* [5], for example, proposed a multi-F0 estimation method based on a deep convolutional neural network (CNN) that estimates an F0 saliency spectrogram from a music spectrogram in the constant-Q transform (CQT) domain, and they applied that method to vocal extraction. Jansson *et al.* [12] used a deep CNN variant with skip connections called a U-Net [17] for estimating a soft mask spectrogram used for separating a vocal spectrogram from a music spectrogram in the short-time Fourier transform (STFT) domain.

This work was supported in part by JST ACCEL Grant Number JPMJAC1602 and JSPS KAKENHI No. JP17K12721 and No. 19H04137.

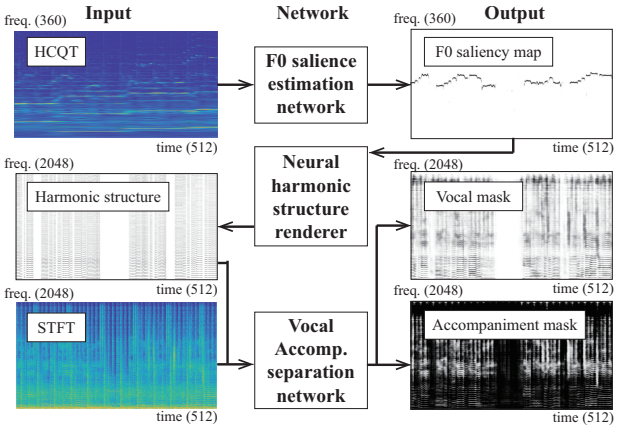

Figure 1: Our multi-task learning architecture consisting of a CNN for vocal extraction and another CNN for vocal separation, between which a neural harmonic structure renderer converts an estimated F0 saliency spectrogram into a harmonic spectrogram in a differentiable manner for guiding vocal separation.

The mutually beneficial relationship between vocal extraction and separation has recently been leveraged for improving the performances of both tasks. Cabañas-Molero *et al.* [18], for example, proposed a three-step method that performs rough vocal separation based on stereo information, autocorrelation-based vocal extraction, and F0-based vocal separation. Hsu *et al.* [19] proposed a tandem algorithm that iterates vocal extraction and separation based on signal processing techniques. To mitigate the error propagation problem of such a cascading approach, Durrieu *et al.* [20] took a machine-learning approach to joint vocal extraction and separation based on source-filter nonnegative matrix factorization (NMF). Mutually beneficial integration of DNN-based vocal extraction and separation, however, has not been achieved yet.

In this paper we propose a unified DNN that effectively combines the deep CNN [5] with the U-Net [12] for joint vocal extraction and separation (Fig. 1). A basic way of connecting these two networks is to warp the frequency scale of an F0 saliency spectrogram estimated by the CNN, stack it on a mixture spectrogram, and feed the two-channel spectrogram into the U-Net. This approach, however, does not incorporate the physical meaning of an F0, *i.e.*, the fundamental knowledge that an F0 indicates an interval between equally spaced harmonic partials, into the unified DNN. An essential research question here is how to parameterize such knowledge

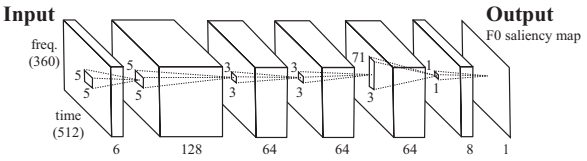

Figure 2: A deep CNN architecture for F0 saliency estimation. The input to each layer is batch normalized. The ReLUs are used as the activation functions of all layers except for the final layer, which uses a sigmoid function to limit the values of F0 saliency to $[0, 1]$.

as a component of the DNN in such a way that the whole DNN can be optimized with standard backpropagation based on the chain rule of the partial derivatives with respect to individual layers.

In the field of speech processing and computer vision, the stable behavior of methods based on fundamental physics has been re-evaluated recently and these methods have been implemented as differentiable layers of DNNs for backpropagation-based supervised training. The neural beamformer [21–23], for example, is a speech enhancement technique that can be combined with DNN-based speech recognition in a jointly trainable manner. Since the classical beamformer is based on linear filtering, it can be implemented as a differentiable layer that converts a noisy speech spectrogram into a clean speech spectrogram by using the spatial information of speech and noise accurately estimated by a DNN. The neural renderer [24] is a rasterization technique that renders 2D images from a 3D polygon mesh in a differentiable manner and can be used for DNN-based 3D reconstruction from 2D images.

Inspired by these studies, we propose a neural harmonic structure renderer that enriches the F0 information estimated by the deep CNN [5] for effectively informing the U-Net [12]. Specifically, an F0 saliency spectrogram is thresholded and converted into a purely harmonic spectrogram indicating the locations of harmonic partials through a differentiable layer that reflects the physical meaning of harmonic structure. The harmonic spectrogram is then used as a scaffold for estimating fine-structured vocal and accompaniment spectrograms thanks to the excellent capability of the U-Net for domain-preserving conversion (*e.g.*, image-to-image conversion). The two DNNs and the parameters of the neural renderer (a salience threshold and weights of harmonic partials) can be optimized jointly with backpropagation through the unified DNN.

The main contribution of this study is to propose a physically founded layer for connecting DNN-based vocal extraction and separation in a differentiable manner. Our study is the first attempt to jointly improve the performances of both tasks by leveraging their mutually relationship in the context of deep learning. We experimentally show that the proposed renderer is reasonably designed.

2. PROPOSED METHOD

We use a deep CNN [5] for F0 saliency estimation (Fig. 2) and use a multi-task extension of the U-Net [12] for vocal and accompaniment separation (Fig. 3). These networks are connected by a neural harmonic structure renderer (Fig. 4).

2.1. Problem specification

Let $\mathbf{X}_{\text{STFT}} \in \mathbb{R}_+^{F\times T}$ and $\mathbf{X}_{\text{HCQT}} \in \mathbb{R}_+^{C\times T\times M}$ respectively be the STFT and harmonic CQT (HCQT) [5] magnitude spectrograms of a music signal, where T, F, C, and M are the number of frames, that of linear frequency bins, that of log-frequency bins, and that of channels. In this paper, $\mathbf{X}_{\text{STFT}}$ and $\mathbf{X}_{\text{HCQT}}$ are computed with the

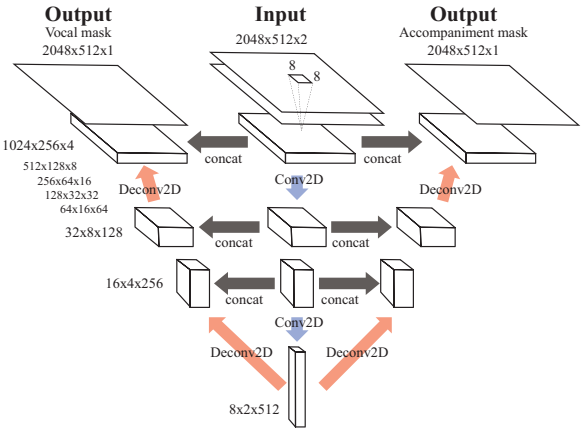

Figure 3: A U-Net architecture for vocal and accompaniment separation. The input to each layer is batch normalized. Leaky ReLUs with leakiness 0.2 are used in the encoder and standard ReLUs are used in the decoder.

librosa library from each channel of a stereo music signal sampled at 44.1 kHz. The STFT is calculated by using a Hann window of 4096 points with a shifting interval of 1024 points ($F = 2048$). The HCQT is computed with 60 bins per octave (20 cents per bin) over six octaves from the lowest frequency of 36.7 Hz ($C = 360$). The M channels of $\mathbf{X}_{\text{HCQT}}$ are obtained by shifting the original CQT of the music signal by -0.5, 0, 1, 2, 3, and 4 octaves ($M = 6$). Both spectrograms are normalized for each song so that the maximum value is 1. For training, the STFT and HCQT spectrograms of a whole musical piece are divided into overlapping segments with a fixed length of $T = 512$ with a shifting interval of 256 samples.

Given $\mathbf{X}_{\text{STFT}}$ and $\mathbf{X}_{\text{HCQT}}$, our goal is to jointly estimate an F0 saliency map $\mathbf{Y}_{\text{F0}} \in \mathbb{R}_+^{C\times T}$ in the log-frequency domain and vocal and accompaniment magnitude spectrograms $\mathbf{Y}_{\text{voc}}$ and $\mathbf{Y}_{\text{acc}} \in \mathbb{R}_+^{F\times T}$ in the linear frequency domain. For vocal extraction (F0 estimation and vocal activity detection (VAD)), we apply a manually specified or automatically learned threshold to $\mathbf{Y}_{\text{F0}}$. For singing voice separation, $\mathbf{Y}_{\text{voc}}$ and $\mathbf{Y}_{\text{acc}}$ are converted to the time-domain signals, by reusing the phase information of $\mathbf{X}_{\text{STFT}}$. For supervised training, we assume that a ground-truth F0 saliency map $\hat{\mathbf{Y}}_{\text{F0}} \in \mathbb{R}_+^{C\times T}$ and ground-truth vocal and accompaniment spectrograms $\hat{\mathbf{Y}}_{\text{voc}}$ and $\hat{\mathbf{Y}}_{\text{acc}} \in \mathbb{R}_+^{F\times T}$ are available.

2.2. Vocal extraction and separation

The vocal extraction network takes $\mathbf{X}_{\text{HCQT}}$ as input and outputs $\mathbf{Y}_{\text{F0}}$ through five convolutional layers (Fig. 2). Based on [5], the first and second layers have 128 and 64 $(5, 5)$ filters, respectively, the following two layers each have 64 $(3, 3)$ filters, and the final layer has 8 $(71, 3)$ filters. At each layer, the stride size is set to 1, and the zero padding is used for keeping the shape of the input. The input to each layer is batch normalized, and each output is passed through a rectified linear unit (ReLU). The final layer uses a sigmoid function to map output of each bin to the range [0, 1].

The vocal separation network takes $\mathbf{X}_{\text{STFT}}$ as input and outputs vocal and accompaniment *mask* spectrograms $\mathbf{M}_{\text{voc}}$ and $\mathbf{M}_{\text{acc}} \in \mathbb{R}_+^{F\times T}$ by using harmonic structure information $\mathbf{Z} \in \mathbb{R}_+^{F\times T}$ (Section 2.3) and calculate $\mathbf{Y}_{\text{voc}}$ and $\mathbf{Y}_{\text{acc}}$ as follows:

$$\mathbf{Y}_{\text{voc}} = \mathbf{M}_{\text{voc}} \odot \mathbf{X}_{\text{STFT}}, \quad \mathbf{Y}_{\text{acc}} = \mathbf{M}_{\text{acc}} \odot \mathbf{X}_{\text{STFT}}, \qquad (1)$$

where $\odot$ indicates the element-wise product, and the vocal and ac-

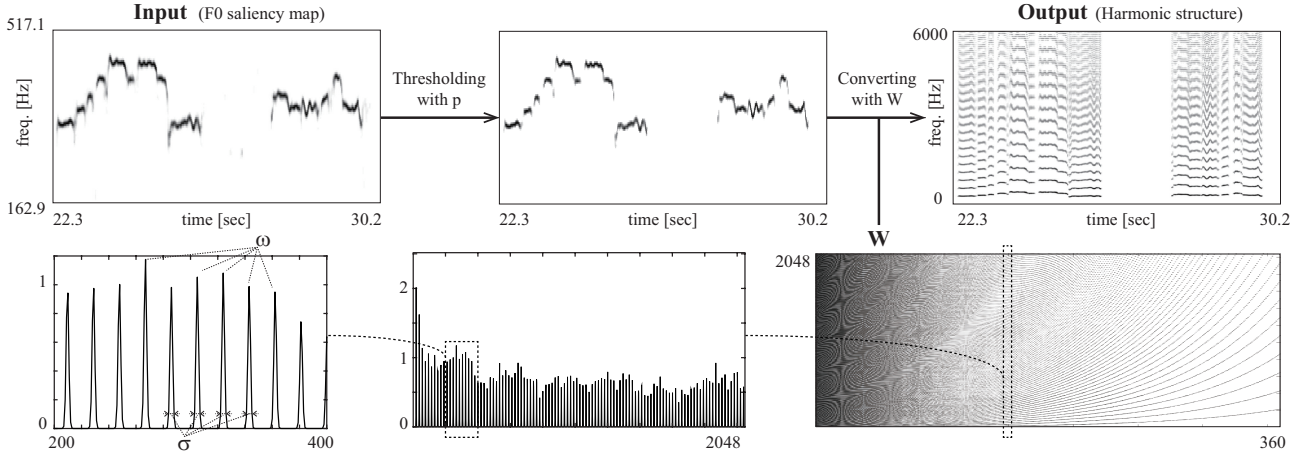

Figure 4: The neural harmonic structure renderer. This network has as parameters p that emphasizes the F0 saliency map and $\boldsymbol{\omega}$ and $\boldsymbol{\sigma}$ that characterize $\mathbf{W}$ to convert the F0 saliency map to a map representing harmonic structure.

companiment masks, $\mathbf{M}_{\text{voc}}$ and $\mathbf{M}_{\text{acc}}$, are soft and in $[0, 1]$.

We extend the U-Net architecture for multi-task learning in which the bottleneck is branched (Fig. 3). We preliminarily found that the extension of adding the decoder for accompaniment separation was effective. Unlike [12], the network takes as input a stack of $\mathbf{X}_{\text{HCQT}}$ and $\mathbf{Z}$ of $(2048, 512, 2)$, which is passed through eight convolutional layers to yield a latent representation of $(8, 2, 512)$ while halving the horizontal and vertical dimensions and doubling the numbers of channels. The filter size, stride size, and zero-padded size are set to 8, 2, and 3, respectively. The output of the encoder is batch normalized and passed through Leaky ReLUs with 0.2 leakiness. The vocal and accompaniment decoders use batch normalization, strided deconvolution with ReLUs, and 50% dropout for each of the first three layers, as in [12].

2.3. Neural harmonic structure renderer

The neural harmonic structure renderer with trainable parameters converts an F0 saliency map $\mathbf{Y}_{\text{F0}} \in \mathbb{R}_+^{C\times T}$ into a harmonic spectrogram $\mathbf{Z} \in \mathbb{R}_+^{F\times T}$ in a differentiable manner by using a dictionary of harmonic spectra $\mathbf{W} \in \mathbb{R}_+^{F\times C}$. This renderer can reflect the physical meaning of an F0, *i.e.*, what the harmonic structure is.

The rendering consists of two steps: thresholding and conversion (Fig. 4). Time-frequency bins with low salience values are ignored by a ReLU with a cut-off value p as follows:

$$\bar{\mathbf{Y}}_{\text{F0}} = \text{ReLU}(\mathbf{Y}_{\text{F0}} - p\mathbf{1}) \odot (\mathbf{1} \oslash (\mathbf{1} - p\mathbf{1})), \tag{2}$$

where $\oslash$ indicates the element-wise division, $\bar{\mathbf{Y}}_{\text{F0}} \in \mathbb{R}_+^{C\times T}$ is a thresholded F0 saliency map, and $\mathbf{1}$ is the all-one matrix. Then $\bar{\mathbf{Y}}_{\text{F0}}$ is converted to $\mathbf{Z}$ as follows:

$$\mathbf{Z} = \mathbf{W}\bar{\mathbf{Y}}_{\text{F0}}, \tag{3}$$

where each column of $\mathbf{W}$ is represented as a weighted sum of Gaussian functions placed on the integral multiples of an F0 as follows:

$$W_{fc} = \sum_{k=1}^{K_c} \omega_k \exp \frac{-(f - k \cdot F_0(c))^2}{2\sigma_k^2}, \tag{4}$$

where ω_k and σ_k^2 are respectively the weight and variance of the k-th Gaussian function, $F_0(c)$ is a function that converts an log-frequency bin index $c \in [1, C]$ (integer) into a linear-frequency bin index (real value), and K_c is set to a maximum value for each c such that $K_c F_0(c)$ does not exceed the Nyquist frequency. p, $\boldsymbol{\omega} = \{\omega_k\}_{k=1}^K$, and $\boldsymbol{\sigma} = \{\sigma_k\}_{k=1}^K$ are the parameters of the renderer that are learned from training data.

2.4. Joint supervised training

The cost function $\mathcal{C}$ of the whole network is the sum of two cost functions $\mathcal{C}_{\text{ext}}$ and $\mathcal{C}_{\text{sep}}$ defined for the vocal extraction and separation networks, respectively, as follows:

$$\mathcal{C} = \mathcal{C}_{\text{ext}} + \mathcal{C}_{\text{sep}}, \tag{5}$$

where

$$\mathcal{C}_{\text{ext}} = \text{CrossEntropy}(\hat{\mathbf{Y}}_{\text{F0}}, \mathbf{Y}_{\text{F0}}), \tag{6}$$

$$\mathcal{C}_{\text{sep}} = |\hat{\mathbf{Y}}_{\text{voc}} - \mathbf{Y}_{\text{voc}}| + |\hat{\mathbf{Y}}_{\text{acc}} - \mathbf{Y}_{\text{acc}}|. \tag{7}$$

Note that $\hat{\mathbf{Y}}_{\text{F0}}$ is obtained as a Gaussian-blurred version of a pure binary salience spectrogram as proposed in [5]. The integrated network is trained with the Adam optimizer [25].

3. EVALUATION

This section reports experiments conducted for evaluating the performances of vocal extraction and separation.

3.1. Experimental conditions

The proposed joint method was compared with the conventional independent methods, *i.e.*, deep CNN (Fig. 2) [5] for vocal extraction and the multi-task learning extension of the U-Net (Fig. 3) [12] for vocal and accompaniment separation. The U-Net takes as input only $\mathbf{X}_{\text{HCQT}}$ of $(2048, 512, 1)$, unlike Fig. 3.

To verify the effectiveness of the neural harmonic structure renderer, three layers that connect the vocal extraction and separation networks were tested.

(1). Fully connected layer (FC): The elements of $\mathbf{W}$ were treated as independent parameters to be optimized instead of using the explicit harmonic parameterization (4).
(2). F0 component renderer (F0): $\mathbf{W}$ was obtained as a dictionary of F0 components by using (4) with $K_c = 1$.

Table 1: Experimental results of joint vocal extraction and separation with different connections of the deep CNN [5] and the U-net [12] with an F0 detection threshold $p_e = 0.1$. The numbers in parentheses are results obtained with a threshold $p_e = p$.

Method	Connection	Updating		Vocal F0 estimation [%]			Vocal activity detection [%]			Separation [dB]	
		$\mathbf{W}$	p	Overall	Raw pitch	Raw chroma	F-measure	Recall	Precision	Vocal	Accomp.
Deep CNN [5]	NA			80.6	79.6	81.4	86.6	92.1	82.7	—	—
U-Net [12]				—	—	—	—	—	—	2.74	9.43
FC-$\mathbf{W}$	Fully connected layer	✓		80.0	80.9	81.9	86.9	93.2	83.0	3.91	9.90
FC-$\mathbf{W}p$		✓	✓	79.3 (77.9)	80.6	81.7	86.2 (85.7)	93.4	81.7	4.12	10.0
F0	F0 component renderer			79.5	80.0	81.1	85.7	92.1	81.5	3.51	9.71
F0-$\mathbf{W}$		✓		80.1	80.6	82.2	86.3	93.2	82.1	3.49	9.45
F0-p			✓	79.5 (55.8)	80.4	81.3	86.0 (75.1)	93.1	80.2	3.72	9.79
F0-$\mathbf{W}p$		✓	✓	79.8 (75.4)	80.2	81.9	86.2 (84.1)	92.8	82.8	3.66	9.63
HS	Harmonic structure renderer			80.4	80.5	81.9	86.8	93.1	83.0	3.98	9.98
HS-$\mathbf{W}$		✓		78.8	80.6	82.3	85.9	93.6	80.0	3.98	10.1
HS-p			✓	**80.9** (82.6)	79.7	81.1	86.8 (87.2)	92.2	**83.7**	3.94	9.91
HS-$\mathbf{W}p$ (ours)		✓	✓	80.5 (82.8)	**83.0**	**84.0**	**87.5** (89.2)	**95.0**	82.9	**4.52**	**10.2**

(3). Harmonic structure renderer (HS): $\mathbf{W}$ was obtained as a dictionary of harmonic patterns by using (4).

The parameters of these layers were treated as follows:

(a). The conversion matrix $\mathbf{W}$ was randomly initialized or its parameters $\boldsymbol{\omega}$ and $\boldsymbol{\sigma}$ were initialized with $\omega_k = 1$ and $\sigma_k = 10$ for every k and then optimized (indicated by "-$\mathbf{W}$"). Otherwise, $\omega = 1$, $\sigma = 10$, and $p = 0.5$ were used.

(b). The threshold p was initialized with $p = 0.5$ and updated (indicated by "-p"). Otherwise, $p = 0.5$ was used.

For evaluation, the MedleyDB dataset [26] and the RWC Music Database [27] were used. The data consisted of 157 songs, 100 from the RWC dataset and 57 from the MedleyDB dataset. We chose the 57 songs that contain vocals with a melody role and for which F0 annotations exist. To evaluate performance by 5-fold cross validation, the 157 songs were randomly divided into three groups of 31 songs and two groups of 32 songs. *mir_eval* [28] was used for computing the source-distortion-ratio (SDR) for vocal separation and several evaluation metrics for vocal extraction. These scores were calculated on each channel for each song, and the average scores over all pieces were then calculated. In vocal F0 estimation, the error tolerance was set to a half semitone (50 cents). To evaluate the performance of VAD, the recall and precision rates and the F-measures were also calculated.

3.2. Experimental results

Table 1 shows the median values of the SDR and the F0 evaluation metrics for the 157 songs. The SDR was computed for both vocal separation and accompaniment separation. To evaluate F0s, we first estimated F0s from an F0 saliency map. At each frame of an F0 saliency map, a frequency that exceeded a threshold p_e and took the maximum value was selected as an F0. In this paper, p_e was set to 0.1. Additionally, for more detailed evaluation, we calculated two representative metrics ("Overall" for F0 estimation and "F-measure" for VAD) with the threshold $p_e = p$. The ground truth F0s were also obtained from the ground truth F0 saliency maps in the same way with $p_e = 0.1$.

With the exception of the "Overall" for F0 estimation and the "Precision" for VAD, results of the proposed method were better than those of all the previous and comparative methods. This suggests that a vocal extraction network and a vocal separation network could be conjoined effectively.

In cross validation the mean optimized p was 0.087 for FC-$\mathbf{W}p$, 0.019 for F0-p, 0.037 for F0-$\mathbf{W}p$, 0.233 for HS-p, and 0.176 for HS-$\mathbf{W}p$. By using the optimized p for the thresholding, better performance could be obtained under the HS-p and the HS-$\mathbf{W}p$, but there were also cases where the performance decreased (HS-$\mathbf{W}p$, F0-p, and F0-$\mathbf{W}p$). These results imply that we could improve our method by designing p optimization to minimize the loss of F0 estimation.

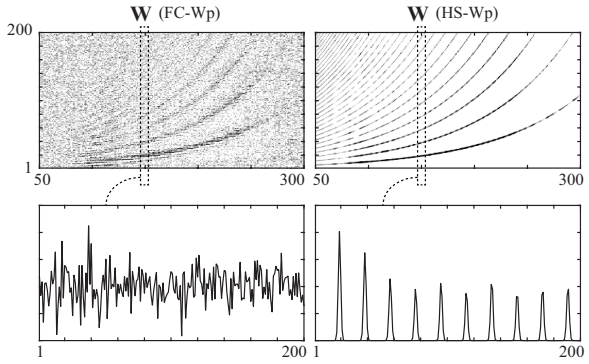

Figure 5: The left figure shows a $\mathbf{W}$ of FC-$\mathbf{W}p$, which is randomly initialized and updated. The right figure shows a $\mathbf{W}$ of HS-$\mathbf{W}p$, updating Gaussian parameters $\boldsymbol{\omega}$ and $\boldsymbol{\sigma}$.

Examples of the optimized conversion matrix $\mathbf{W}$s for FC-$\mathbf{W}p$ and HS-$\mathbf{W}p$ are shown in Figure 5. The FC-$\mathbf{W}p$ acquired a conversion to harmonic structure automatically, but much noise remains. The harmonic structure is obtained only in the range where the vocal F0 is frequently observed and in the lower harmonic components. In contrast, $\mathbf{W}$ in HS-$\mathbf{W}p$ represents clean harmonic structures, which may have contributed to the performance improvement.

4. CONCLUSION

In this paper we propose a neural network architecture that jointly performs vocal extraction and separation for polyphonic music signals. We confirmed its effectiveness experimentally. All parameters of the proposed network connected by a differentiable renderer can be optimized simultaneously by backpropagation.

While the ground-truth F0s and singing and accompaniment signals were used for completely supervised training in our experiment, semi-supervised training would in theory be feasible when some ground-truth data are missing. We also plan to deal with unvoiced utterances and make effective use of phase information to reduce the distortion of a separated singing voice.

5. REFERENCES

[1] A. Demetriou, A. Jansson, A. Kumar, and R. M. Bittner, "Vocals in music matter: the relevance of vocals in the minds of listeners," in *ISMIR*, 2018, pp. 514–520.

[2] J. Salamon, E. Gómez, D. P. W. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[3] Z. Rafii, A. Liutkus, F.-R. Stoter, S. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Transactions on Audio, Speech, Language Processing*, vol. 26, no. 8, pp. 1307–1335, 2018.

[4] F. Rigaud and M. Radenen, "Singing voice melody transcription using deep neural networks," in *ISMIR*, 2016, pp. 737–743.

[5] R. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for F0 estimation in polyphonic music," in *ISMIR*, 2017, pp. 63–70.

[6] D. Basaran, S. Essid, and G. Peeters, "Main melody extraction with source-filter nmf and crnn," in *ISMIR*, 2018, pp. 82–89.

[7] W.-T. Lu and L. Su, "Vocal melody extraction with semantic segmentation and audio-symbolic domain transfer learning," in *ISMIR*, 2018, pp. 521–528.

[8] 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.

[9] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *ICASSP*, 2015, pp. 2135–2139.

[10] Z. C. Fan, J. S. R. Jang, and C. L. Lu, "Singing voice separation and pitch extraction from monaural polyphonic audio music via DNN and adaptive pitch tracking," in *BigMM*, 2016, pp. 178–185.

[11] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *ICASSP*, 2017, pp. 61–65.

[12] A. Jansson, E. J. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *ISMIR*, 2017, pp. 745–751.

[13] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *ISMIR*, 2017, pp. 330–340.

[14] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band DenseNets for audio source separation," in *WASPAA*, 2017, pp. 21–25.

[15] K. W. E. Lin, B. T. Balamurali, E. Koh, S. Lui, and D. Herremans, "Singing voice separation using a deep convolutional neural network trained by ideal binary mask and cross entropy," *Neural Computing and Applications*, pp. 1–14, 2018.

[16] D. Stoller, S. Ewert, and S. Dixon, "Jointly detecting and separating singing voice: A multi-task approach," in *LVA/ICA*, 2018, pp. 329–339.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[18] P. Cabañas-Molero, D. M. Muñoz, M. Cobos, and J. J. López, "Singing voice separation from stereo recordings using spatial clues and robust F0 estimation," in *AEC Conference*, 2011.

[19] C. L. Hsu, D. Wang, J. R. Jang, and K. Hu, "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 20, no. 5, pp. 1482–1491, 2012.

[20] J. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1180–1191, 2011.

[21] T. N. Sainath, A. Narayanan, R. J. Weiss, E. Variani, K. W. Wilson, M. Bacchiani, and I. Shafran, "Reducing the computational complexity of multimicrophone acoustic models with integrated feature extraction," in *Interspeech*, 2016, pp. 1971–1975.

[22] B. Li, T. N. Sainath, R. J. Weiss, and K. W. Wilson, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Interspeech*, 2016, pp. 1976–1980.

[23] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[24] H. Kato, Y. Ushiku, and T. Harada, "Neural 3D mesh renderer," in *CVPR*, 2018, pp. 3907–3916.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015, pp. 1–15.

[26] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 155–160.

[27] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2002, pp. 287–288.

[28] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, , and D. P. W. Ellis, "mir_eval: A transparent implementation of common mir metrics," in *ISMIR*, 2014, pp. 367–372.