Weakly informed audio source separation

Kilian Schulze-Forster, Clément Doire, Gael Richard, Roland Badeau

WEAKLY INFORMED AUDIO SOURCE SEPARATION

Kilian Schulze-Forster,[1*] *Clément Doire,*[2] *Gaël Richard,*[1] *Roland Badeau*[1]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
firstname.lastname@telecom-paris.fr
[2] Audionamix, 171 quai de Valmy, 75010 Paris, France

ABSTRACT

Prior information about the target source can improve audio source separation quality but is usually not available with the necessary level of audio alignment. This has limited its usability in the past. We propose a separation model that can nevertheless exploit such weak information for the separation task while aligning it on the mixture as a byproduct using an attention mechanism. We demonstrate the capabilities of the model on a singing voice separation task exploiting artificial side information with different levels of expressiveness. Moreover, we highlight an issue with the common separation quality assessment procedure regarding parts where targets or predictions are silent and refine a previous contribution for a more complete evaluation.

Index Terms— informed source separation, singing voice separation, weak labels, attention, separation evaluation

1. INTRODUCTION

Recent deep learning based methods for audio source separation are trained in a supervised fashion on mixture and target source pairs [1, 2, 3]. Even though they achieve good separation results on test sets such as MUSDB18 [4] for singing voice separation, they are not flawless. The best performing algorithms in the Signal Separation Evaluation Campaign (SiSEC) 2018 [5], i.e. TAK, TAU, UHL, achieve Source-to-Distortion Ratio (SDR) [6] scores in the range $[-14.7; 17]$ dB on the 50 test songs [7]. This shows that training a model that generalizes well to all music styles is a difficult task, even with large and diverse training data.

As opposed to these purely data-driven methods, the informed source separation approach exploits prior information about the target source [8], making systems more adaptive to observed signals. It has been shown that source separation can benefit from side information in the form of a musical score [9], the target source pitch [10], a text transcript [11], or visual clues [12], among others. Recently, data-driven and informed approaches have been combined [13, 14, 15].

However, the main obstacle for exploiting side information remains: accurately labeled data is expensive to create and thus rare. For example, aligning a score on the note level or lyrics on the phoneme level would require manual annotations, as creating such fine alignment automatically remains an open problem [15, 16]. On the other hand, weak side information such as non-aligned scores or lyrics is often easily available but not straightforward to employ. Consequently, some form of automatic alignment is usually applied before the actual separation and the usefulness of the side information then depends on the quality of such an alignment [11, 13, 14, 15].

Seeking to combine the power of data-driven models with the adaptability of informed approaches we propose a deep learning based separation method that employs very weak side information with extremely coarse alignment. It is based on the attention mechanism, which was proposed to learn an alignment between two sequences while performing another task such as machine translation [17, 18]. Attention became a widely adopted concept and has, for example, been used for automatic speech recognition [19] and image generation [20]. In [21] image captions are translated taking different parts of the image content as additional side information into account. Inspired by this approach, we propose a source separation model with a sequential encoder-decoder architecture where the decoder is connected to the side information via attention. The whole side information sequence is thus accessible to the decoder at all time steps. During training, it learns to evaluate the relevance of the side information elements with respect to the separation task. This relevance is reflected in attention weights from which alignment information can be retrieved.

Training models with weakly labeled data remains a challenging problem for a variety of audio related tasks [14, 15, 22, 23]. In this context, Multi-Instance learning (MIL) has been applied to singing voice detection [22] and acoustic event detection [23] to gradually refine the labeling during supervised training, but with limited effectiveness [14, 22]. For informed source separation, it has been proposed to approach training with weak labels in an unsupervised fashion [14]. The side information is then used to enforce structure on the latent representation of the mixture within an autoencoder model. While the approaches above aim for training with weakly labeled data only, we intend to complement supervised strong label training with additional weaker information. In [15] a tolerance window allows for misalignment of around 0.2 seconds during score-informed source separation. Instead of explicitly guiding the network regarding how to use the side information as in [14, 15], we let our model learn the best use for the separation task and allow for even weaker side information.

It has been tested in [11] if the alignment of side information can be improved during text-informed source separation. No improvement over the pre-alignment could be reported, while the authors stated that it would have been beneficial for the separation quality. We show in experiments that our model can indeed improve the alignment by a considerable extent.

In short, our contributions are the following: adapting the attention mechanism to informed singing voice separation and thereby allowing the use of very weak side information, learning an alignment as byproduct. Moreover, we highlight an issue with the common source separation evaluation procedure regarding silent signal parts and refine the solution previously suggested in [24].

*This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowsa-Curie grant agreement No. 765068.

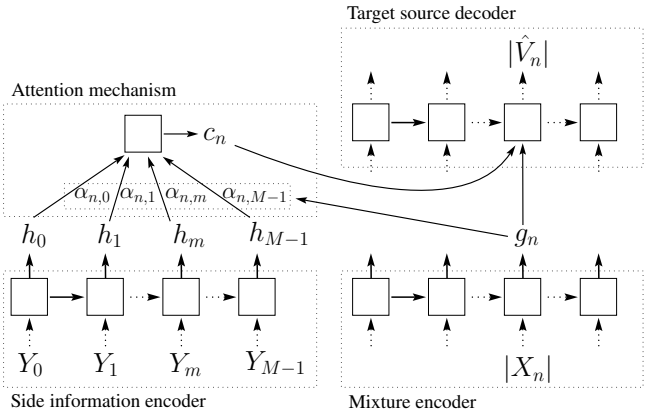


Figure 1: Schematic model architecture and workflow of the attention mechanism to compute prediction frame $|\hat{V}_n|$.

2. PROPOSED MODEL

Let $x(t)$ be the observed single-channel mixture signal at discrete-time t. Our goal is to separate $x(t)$ into a target source $v(t)$ and a mixture of all remaining sources $a(t)$. Let $Y \in \mathbb{R}^{D\times M}$ be a side information sequence with feature dimension D and M time steps.

The proposed model takes as inputs the magnitude of the mixture's Short Time Fourier Transform (STFT) $|X| \in \mathbb{R}^{F\times N}$ with F frequency bands and N time frames as well as the information Y. The output is an estimate of the target's magnitude STFT $|\hat{V}| \in \mathbb{R}^{F\times N}$. An inverse STFT of $|\hat{V}|$ combined with the mixture phase is performed to obtain the target estimation $\hat{v}(t)$ in the time domain. Assuming a linear mixture model, the estimation of remaining sources $\hat{a}(t)$ is obtained as $\hat{a}(t) = x(t) - \hat{v}(t)$.

2.1. Architecture details

The proposed model comprises four building blocks, namely a mixture encoder, a side information encoder, an attention mechanism, and a target source decoder as shown in Figure 1.

The mixture encoder is a two-layer deep Bidirectional Recurrent Neural Network (BRNN) [25] with Long Short-Term Memory (LSTM) cells [26]. Given the sequence of mixture STFT time frames $|X_n|$ with $n = 0, ..., N-1$, it computes the sequence $g \in \mathbb{R}^{E\times N}$, which we call the mixture encoding. It has feature dimension E and length N over time. Variables indexed by n and m refer to the n-th and m-th sequence element, respectively.

The side information encoder has the same architecture as the mixture encoder. Given the sequence of side information frames Y_m with $m = 0, ..., M-1$, it computes the encoding of the side information $h \in \mathbb{R}^{J\times M}$ with feature dimension J.

The target source decoder gets as inputs the mixture encoding g and a representation of the side information encoding denoted c, which is computed by the attention mechanism as explained below. Both inputs are concatenated along the feature dimension, which is denoted by $[c_n, g_n]$. The decoder computes one time frame of the target source estimation $|\hat{V}_n|$ through the following layers. W and b are learnable weights and biases respectively in the equations below. First, a fully connected layer computes the hidden representation $q_n^{(1)}$:

$$q_n^{(1)} = \tanh(W_1[c_n, g_n] + b_1). \quad (1)$$

Then, a two layers deep BRNN with LSTM cells – just as in the encoders – computes the hidden representation $q_n^{(2)}$. Finally, another fully connected layer with ReLU activation computes the estimation:

$$|\hat{V}_n| = \max(0, W_2 q_n^{(2)} + b_2). \quad (2)$$

Predicting time-frequency masks as in [27] instead of magnitude spectrograms directly did not lead to better results in our experiments.

The attention mechanism identifies the relevant elements in the side information sequence for each time step n of the target source decoding and summarizes them in a context vector c_n. Consequently, the decoder can find at every time step the relevant side information elements no matter where they are placed in the sequence, which makes a pre-alignment redundant. We closely follow the attention mechanism proposed in [17] and refined in [18].

For time step n of the decoder, the vector c_n is computed as follows. A score $s_{n,m}$ is calculated representing some similarity or "energy" between the mixture encoding time step g_n and each of the side information encoding steps h_m:

$$s_{n,m} = g_n^\top W_s h_m \ \ \forall m \in \{0, 1, ...M-1\} \quad (3)$$

where $W_s \in \mathbb{R}^{E\times J}$ is a matrix of learnable weights.

Then, attention weights $\alpha_{n,m}$ are computed from the scores by a softmax operation:

$$\alpha_{n,m} = \frac{\exp(s_{n,m})}{\sum_{m=0}^{M-1} \exp(s_{n,m})}. \quad (4)$$

Each element in the side information h_m thus has a dedicated weight $\alpha_{n,m}$ reflecting its importance for the decoder time step n as a probability. The context vector c_n is the weighted sum of all side information encoding elements:

$$c_n = \sum_{m=0}^{M-1} h_m \alpha_{n,m}. \quad (5)$$

The target source estimation is then computed from the context vector and the mixture encoding g_n as described above. The alignment between mixture and side information is reflected in the attention weights $\alpha_{n,m}$ and is learned without any additional term in the loss function.

We use BRNNs because they treat data sequentially, which makes the application of attention more straightforward and easier to illustrate. However, with some modifications, the attention mechanism could also be applied to Convolutional Neural Network (CNN) architectures [20].

3. EVALUATION

The most commonly used metrics for source separation performance evaluation are Source-to-Distortion Ratio (SDR), Source-to-Artifacts Ratio (SAR), and Source-to-Interference Ratio (SIR) [6]. They are typically computed on non-overlapping frames of one second length and the median is taken to represent the performance on the whole signal [5]. However, for frames with a silent true source or prediction, the metrics are undefined [6]. The MUSDB test set [4] has 2600 such frames with silent vocals and 103 frames with silent accompaniment. As a result, at least about 45 out of 210 minutes are systematically ignored during evaluation, with potentially more frames being ignored when the prediction is silent. This issue

has also been observed in [24], where the authors suggest reporting the root mean square energy of the prediction for frames with silent ground truth. Following the suggestion, we report the Predicted Energy at Silence (PES) score for each test song. It is the mean of the energy in the predictions at those frames with silent ground truth. It reflects a method's capability to not get confused by other sources while the target is not active.

However, in order to include every single test frame in the evaluation, we also need to evaluate frames for which silence is predicted while the ground truth is not silent. Therefore, we propose to report also the Energy at Predicted Silence (EPS) score, which is the mean of the ground truth energy of all frames with silent prediction and non-silent ground truth. Frames with silent ground truth are already included in the PES. The EPS reflects a method's capability to predict silence at the correct time. It makes the implicit assumption that musically meaningful signal parts have more energy than background noise.

4. EXPERIMENTS

We perform monaural singing voice separation with the proposed model using artificial side information about the singing voice with different levels of expressiveness.

4.1. Data sets

We use the publicly available data set MUSDB18 [4] comprising a 100 tracks training set and a 50 tracks test set containing various genres. We split the training set into 80 tracks for training and 20 tracks for the validation set. All songs are converted to mono, downsampled to 16 kHz, and cut into fragments of 8.2 seconds. The STFT is computed on each fragment with Fast Fourier Transform (FFT) length 1024, Hamming window, and hop length of 512 leading to magnitude spectrograms of size $(F \times N) = (513 \times 256)$. Each magnitude spectrogram is divided by its maximum value to normalize it to the range $[0; 1]$.

As data augmentation we set the energy ratio between vocals and accompaniment to a value uniformly drawn from the ± 2 dB range around the original energy ratio. We also shift the mixture's pitch by w half tone steps, with w being uniformly drawn from $[-2; 2]$. These random operations are repeated four times on each original fragment leading to 8152 fragments for training in total.

We use this limited amount of publicly available data to make our results easier to reproduce for fellow researchers. However, it is not straightforward to evaluate whether performance of data-driven methods is limited by the model's architecture or the amount of training data [5]. We therefore repeat all experiments with additional training data (65 rock-pop song excerpts with 96 minutes total length) to test if performance is scalable.

4.2. Training

We train the model on batches of 128 spectrograms randomly drawn from the training set. The loss function is the L1 loss. The ADAM optimizer [28] is used with learning rate 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay rate 0.001. We set both the size E of the mixture encoding and size J of the side information encoding to 513. We select the model with the lowest validation cost after 100 epochs without improvement of the validation cost.

4.3. Side information

The side information Y has length M, which can be equal to or different from the mixture length N. We use side information with feature size $D = 1$.

We use two baseline models. The first one (BL1) only consists of the mixture encoder and target decoder. It does not use any side information. As second baseline (BL2) we use the full proposed architecture, which is also used in all subsequent experiments, and provide only meaningless side information: a sequence of ones. This allows us to investigate to which extent the added learning capacity of the attention mechanism and side information encoder improves performance. Next, we investigate whether performance can be further improved with meaningful side information.

First, we provide the total vocals Magnitude (M) for each time frame as side information. It is derived from the ground truth spectrograms by summing the magnitudes of all frequencies at each time step n: $Y = \sum_{f=0}^{F-1} |V_{f,n}|$. It is considered as very strong information, since it has the same length as the mixture ($M = N = 256$) and is numerically closely related to the ground truth. We call it M1. We then derive M2 from it by padding both sides of the sequence so that $M = 300$. We use 100 as padding value and randomly choose the padding length on both sides for each batch. As a result, M2 conveys the same strong information as M1 but is less synchronized to the mixture. The position of relevant information varies from batch to batch during training and from example to example during testing.

Binary sequences indicating vocal Activity (A) (1) and non-activity (0) are derived from M1 by setting all time steps m with total magnitude values below 0.1 to 0 and all other steps to 1. In practice, such weak information can be obtained by vocal activity detection methods [29]. For experiment A1 we pad the binary sequence to length $M = 300$ keeping the padding value 100 following the procedure of M2 to de-synchronize it from the mixture.

For experiment A2, we further weaken the information by deleting a random number w of zeros in each sub-sequence of zeros in the binary sequence. We draw w uniformly from $[1; L/2]$ for each example, where L is the length of a sub-sequence of zeros. We pad the remaining binary sequence to length $M = 300$ as above. The sequence Y now only contains information about the number of appearances of silent parts in the vocals and their position relative to non-silent parts. Information about the silence length is almost completely lost.

For experiment A3, we weaken the binary information even further by additionally reducing the length of sub-sequences of ones with the same rule as applied to zeros in A2. We also pad to length $M = 300$. Now, Y carries only information about the alternations between vocal activity and silence. We test the model trained with this side information in two different inference settings. First, with the same side-information as seen during training (A3.1), then with this side information circularly shifted by 100 steps (A3.2).

5. RESULTS AND DISCUSSION

The evaluation results are shown in Figure 2. Due to space constraints we only show a limited number of results for the case of additional training data indicated by '+'. The results of all experiments with additional data are available online (https://schufo.github.io/publication/2019-WASPAA), as well as audio examples and a PyTorch implementation of the proposed model. The relative results do not change substantially when us-

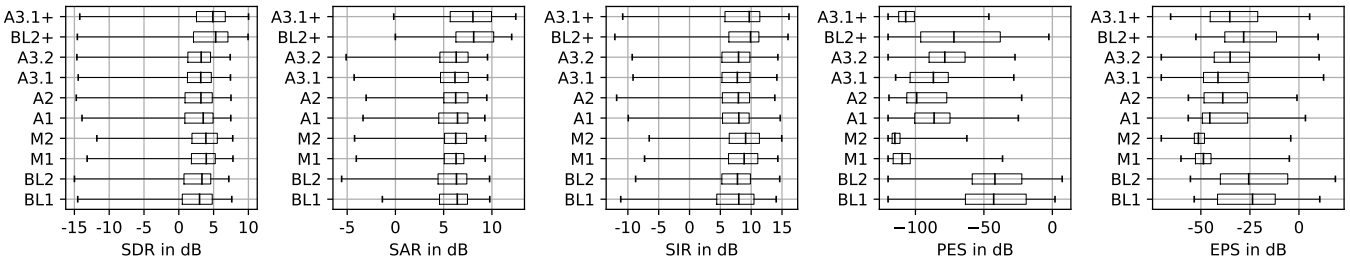
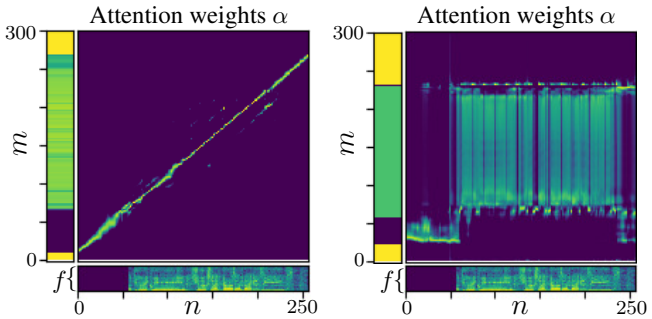


Figure 2: Source separation evaluation results. For SDR, SAR, SIR higher values are better, while for PES and EPS lower values are better. BL: baseline, M: vocal magnitude side information, A: vocal activity side information. The '+' indicates use of additional training data.



(a) Experiment M2 (b) Experiment A3.1

Figure 3: Attention weights α containing alignment information. The side information is shown vertically on the left of α and the true vocals spectrogram below. Lighter color indicates higher values.

ing more training data. Each data point represents the median over all evaluation frames of one test song following the procedure described in Section 3. The box extends from lower to upper quartile with the line inside representing the median. The whiskers extend over the whole data range. Note that for the proposed PES and EPS metric lower values are better, while for the standard metrics higher values are better.

The baselines BL1 and BL2 achieve a median SDR of 3.0 dB and 3.33 dB respectively, which, given the amount of training data and simplicity of the model, can be considered an appropriate baseline. The improvement of BL2 over BL1 shows that the proposed model can leverage the additional capacity even with meaningless side information. Adding only 96 minutes of training data (BL2+) improves performance on all metrics so that the baseline would have only been outperformed by models trained on much more data in the SiSEC 2018 [5].

The use of all types of meaningful side information considerably improves a) performance on silent vocal frames resulting in a much lower PES and b) predicting silence at the right time resulting in a lower EPS. In case of M1 and M2, the SDR and SIR are also improved, while with the binary vocal activity side information the standard metrics do not change much compared to the baselines. These observations are in line with [24]. For frames with high vocal energy, a lot of information about the vocals is already contained in the mixture. Consequently, the binary side information does not add information for these frames, while the vocal magnitude information does. For frames with silent or near-silent vocals, any other source can potentially be mistaken as vocals leading to wrong predictions. In this case the binary information is useful to understand the alternations between vocal activity and non-activity. The fact that M2 performs slightly better than M1 can be explained by the data augmentation effect of the random padding in M2.

In general, it is not surprising that additional information leads to better separation results. Our contribution lies rather in the fact that the proposed model can exploit such information despite its weakness. Note that the binary side information types carry less information than a musical score.

In addition to improving source separation performance by exploiting weak side information, the proposed model also provides an alignment estimation between the side information and the mixture through the attention weights. In Figure 3 the attention weights α are shown for experiments M2 and A3.1 for one fragment of the MUSDB18 test track *Schoolboy Fascination*, which is also available as audio example. On the left of each matrix α the corresponding side information is depicted vertically with time step m. Dark blue indicates a zero value, while padding is shown in yellow. Below α the true vocals spectrogram is shown with frequency bands f and time frames n. The lighter the color at point (n, m) the more the side information element at m is taken into account for producing the prediction at time step n. For M2 a very exact alignment to the mixture is learned, it becomes a bit blurry at the silent vocal part, where the side information contains low and therefore similar values. For A3.1 the model learned to look at ones and zeros at the right time, although the sub-sequences are much shorter than the corresponding parts in the true vocals. The model learned to never look at the padding values. The attention weights α show that the model has indeed learned to find the relevant side information at each time step without any pre-alignment.

6. CONCLUSION

In this paper, we proposed a model that includes weak side information via attention during audio source separation. We demonstrated its capability not only to exploit weak side information but also to align it on the mixture as a byproduct on a singing voice separation task with artificial side information. This can increase the usability of side information such as scores or lyrics transcripts, that previously suffered from inaccurate pre-alignments. Moreover, we refined a previous solution regarding separation quality evaluation for signal frames with a silent target or prediction in order to enable assessment of the entire signal. In the future, we plan to extend our work to scores and lyrics as side information and to evaluate the alignment estimation more thoroughly.

7. REFERENCES

[1] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 23–27, 2017.

[2] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 21–25, 2017.

[3] S. Park, T. Kim, K. Lee, and N. Kwak, "Music source separation using stacked hourglass networks," *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 289–296, 2018.

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

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

[6] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[7] "Sisec mus 2018 objective evaluation results." https://sisec18.unmix.app/#/results/vocals/SDR, accessed: 2019-04-18.

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

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

[10] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music." *INTERSPEECH*, pp. 17–22, 2008.

[11] L. Le Magoarou, A. Ozerov, and N. Q. Duong, "Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 117–131, 2015.

[12] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audio-visual speech source separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125–134, 2014.

[13] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] S. Ewert and M. B. Sandler, "Structured dropout for weak label and multi-instance learning and its application to score-informed source separation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2277–2281, 2017.

[15] M. Miron, J. Janer Mestres, and E. Gómez Gutiérrez, "Monaural score-informed source separation for classical music using convolutional neural networks," *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 55–62, 2017.

[16] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," *arXiv preprint arXiv:1902.06797*, 2019.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[18] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[19] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4960–4964, 2016.

[20] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[21] P.-Y. Huang, F. Liu, S.-R. Shiang, J. Oh, and C. Dyer, "Attention-based multimodal neural machine translation," *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, vol. 2, pp. 639–645, 2016.

[22] J. Schlüter, "Learning to pinpoint singing voice from weakly labeled examples." *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 44–50, 2016.

[23] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1038–1047, 2016.

[24] D. Stoller, S. Ewert, and S. Dixon, "Jointly detecting and separating singing voice: A multi-task approach," *International Conference on Latent Variable Analysis and Signal Separation*, pp. 329–339, 2018.

[25] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] S. I. Mimilakis, K. Drossos, J. F. Santos, G. Schuller, T. Virtanen, and Y. Bengio, "Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 721–725, 2018.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks." *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 121–126, 2015.