End-to-end music source separation: is it possible in the waveform domain?

Francesc Lluis*, Jordi Pons*, Xavier Serra

Music Technology Group, Universitat Pompeu Fabra, Barcelona.

Index Terms: source separation, end-to-end learning.

1. Introduction

When two or more sounds co-exist, they interfere with each other resulting in a novel mixture signal where sounds are superposed (and, sometimes, masked). The source separation task tackles the inverse problem of recovering each individual sound source contribution from an observed mixture signal.

With the recent advances in deep learning, source separation techniques have improved substantially [1]. Interestingly, though, nearly all successful deep learning algorithms use the magnitude spectrogram as input [1, 2, 3] — and are therefore, by default, omitting part of the signal: the phase. Omitting the potentially useful information of the phase entails the risk of finding a sub-optimal solution. In this work, we aim to take full advantage of the acoustic modeling capabilities of deep learning to investigate whether it is possible to approach the problem of music source separation directly in the end-to-end learning fashion. Consequently, our investigation is centered on studying how to separate music sources (e.g., singing voice, bass or drums) directly from the raw waveform mixture.

During the last two decades, matrix decomposition methods have dominated the field of audio source separation. Several algorithms have been proposed throughout the years, with independent component analysis (ICA) [4], sparse coding [5], or non-negative matrix factorization (NMF) [6] being the most used ones. Given that magnitude or power spectrogram representations are always non-negative, imposing a non-negative constraint (like in NMF) is particularly useful when analyzing these spectrograms — but less appropriate for processing waveforms, which range from -1 to 1. For that reason, methods like ICA and sparse coding have historically been used to process waveforms [7, 8, 9]. Waveform representations preserve all the information available in the raw signal. However, given the unpredictable behavior of the phase in real-life sounds, it is rare to find identical waveforms produced by the same sound source. As a result of this variability, a single basis cannot represent a sound source and therefore, one requires i) a large amount of bases, or ii) shift-invariant bases to obtain accurate decompositions [8, 10]. Although several matrix decomposition methods have been used for decomposing waveform-based mixtures [7, 8, 9], these have never worked as well as the spectrogram-based ones.

Due to the above mentioned difficulties, the phase of complex time-frequency representations is commonly discarded, assuming that magnitude spectrograms already carry meaningful information about the sound sources to be separated. Phase related problems disappear when sounds are just represented as magnitude or power spectrograms, since different realizations of the same sound are almost identical in this time-frequency plane. This allows to easily overcome the variability problem found when operating with waveforms.

Most matrix decomposition methods rely on a signal model assuming that sources add linearly in the time domain [10].

However, the addition of signals in the time and frequency domains is not equivalent if phases are discarded. Only in expectation: $E[|X(k)|^2] = |Y_1(k)|^2 + |Y_2(k)|^2$, where $X(k) = DFT\{x(t)\}$. This means that we can approximate the time-domain summation in the power spectral domain. For that reason, many approaches utilize power spectrograms as inputs. Although magnitude spectrograms work well in practice [11], there is no similar theoretical justification for such an inconsistency with the signal model when the phases are discarded.

Finally, note that these methods operating on top of spectrograms still need to deliver a waveform signal. To this end, the main practice is to filter the original magnitude or power spectrogram with (predicted) time-frequency masks. Accordingly, the original noisy phase of the mixture is used when synthesizing the waveform of the estimated sources — which might introduce an additional source of error [10]. Notably, many modern spectrogram-based deep learning models are also relying on this same (potentially problematic) approach [2, 12].

To overcome this issue, some tried to consider the phase when separating the sources [13, 14, 15], or some others relied on a sinusoidal signal model at synthesis time [16]. However, in our work, we do not want to rely on any time-frequency transform or any signal model. Instead, we aim to directly approach the problem in the waveform domain.

As seen, many issues still exist around the idea of discarding the phase: are we missing crucial information when discarding it? When using the phase of the mixture at synthesis time, are we introducing artifacts that are limiting our model’s performance? Or, since magnitude spectrograms (differently from

---

*Contributed equally.
Our goal is to address these historical challenges by bypassing the problem. We want to investigate the feasibility of counting on an end-to-end model instead of relying on any signal model, any time-frequency transform, or filtering any signal. However, waveforms are high-dimensional and very variable. Thus, is music source separation possible in the waveform domain? Recent literature shows that deep learning models operating on raw audio waveforms can achieve satisfactory results for several audio-based tasks [17, 18, 19]. And, among those, some are also recently starting to address the problem of music source separation directly in the waveform domain [20, 21].

To this end, we set Wave-U-Net as one of our baseline and run a perceptual study to get a broader picture of how end-to-end learning models can perform. As seen, the idea of approaching the music source separation task directly in the waveform domain has not been widely explored throughout the years, possibly due to the complexity of dealing with waveforms (which are unintuitive and high-dimensional). Consequently, during the last decades, music source separation in the waveform domain has been considered almost unattainable. Our work aims to keep adding knowledge on top of this rather scarce literature, to convince the reader that addressing music source separation in an end-to-end learning fashion is possible. To this end, we set Wave-U-Net as one of our baseline models and we adapt it for monaural music source separation. Figure 1 shows an overall depiction of the model, where we can observe that every layer has residual and skip connections. Before anything else, the waveform is linearly projected to a tensor (k channels by a 3x1 CNN layer) to comply with the feature map dimensions of each residual layer. Then, this projection is processed with several residual and skip connections. A ReLU is applied after summing all resulting feature map dimensions to be the same as the residual dimension. Consequently, during the last decades, music source separation in the waveform domain has been considered almost unattainable. Our work aims to keep adding knowledge on top of this rather scarce literature, to convince the reader that addressing music source separation is possible in the waveform domain. To this end, we set Wave-U-Net as one of our baseline models for music source separation. Later, we present two recent deep learning models that we set as baseline for our study: DeepConvSep and Wave-U-Net.

In sections 3 and 4 we evaluate the above mentioned models, to conclude in section 5 that performing music source separation directly in the waveform domain is not only possible, but it can be a promising research direction. Hence, our main contributions can be summarized as follows:

1) We propose to use a Wavenet-based model for music source separation. Besides, we study the impact of several hyper-parameters — a result that might also be of relevance for other application areas where Wavenet has been used.

2) We perceptually benchmark several music source separation models, including our Wavenet-based model. This first perceptual study helps to further understand which might be the contribution of end-to-end models to the field of source separation.

2. End-to-end source separation models

We aim to discuss the feasibility of end-to-end learning models for monaural music source separation. To this end, we experiment with a new Wavenet-based model for music source separation we propose, and we compare it against two recent models: DeepConvSep, a spectrogram-based deep learning model for multi-instrument separation; and Wave-U-Net, a waveform-based model trained end-to-end for singing voice separation. We will compare the performance of these models perceptually and via assessing their BSS Eval scores. For a fair comparison, all discussed models are trained with MUSDB data down-sampled at 16 kHz.

2.1. A Wavenet-based model for source separation

We utilize an adaptation of Wavenet [24] that turns the original causal Wavenet (that is generative and slow), into a non-causal model (that is discriminative and parallelizable). This idea was originally proposed by Rethage et al. [25] for speech denoising, and we adapt it for monaural music source separation. Figure 1 shows an overview of the non-causal Wavenet we propose for multi-instrument source separation.

![Figure 1: Left – Residual layer. Right – Overview of the non-causal Wavenet we propose for multi-instrument source separation.](image)

3It is multi-resolution in the sense that they use several CNN filter lengths at every layer so that short- and long-term features can be efficiently learned/encoded. For further information, see Pons et al. [18].
has 3 outputs; and for singing voice separation, it has 1 single output. The remaining sources are computed via subtracting the estimated sources from the mixture, see Figure 1 (Right). The main difference between the original Wavenet and the non-causal adaptation we use, is that some samples from the future can be used to predict the present one. As a result of removing the autoregressive causal nature of the original Wavenet, this fully convolutional model is able to predict a target field instead of one sample at a time — due to this parallelization, it is possible to run the model in real-time on a GPU [25]. Another major difference with the original Wavenet is the output: we directly regress the waveform sources instead of sampling from a softmax output [25]. We minimize the mean absolute error (MAE) regression loss during training. ADAM optimizer is used with a learning rate of 0.001. We set the batch size to 10, and the model is trained until the validation error does not decrease for 16 epochs. The model with the lowest validation loss is selected. The code is accessible online.6

2.2. DeepConvSep: a spectrogram-based model

DeepConvSep [2] is a state-of-the-art spectrogram-based model that is openly available. Following the common practice: mixture signals (pre-processed as magnitude spectrograms) are fed to the model to estimate time-frequency soft masks for each source [1, 3, 12]. These masks are then used to filter the magnitude spectrogram of the mixture to estimate the magnitude spectrograms of the separated sources. Finally, these estimates, along with the phase of the mixture, are used to obtain the waveform signals corresponding to the separated sources. DeepConvSep’s architecture is based on a convolutional encoder-decoder. The encoder is composed by a first CNN layer with 50 vertical filters aiming to capture timbral representations [26], a second CNN layer with 30 horizontal filters modeling temporal cues [27], and a dense layer with 128 units acting as a bottleneck. The decoder contains two deconvolutional layers which up-sample the bottleneck feature maps up to have the same input size, which correspond to the estimated masks. The model learns via minimizing the mean squared error (MSE), together with several dissimilarity loss terms [2, 12]. Which cost? Table 1 describes the setups we study.

Table 1: Description of the models we study. Wave-U-Net-based “k filters” stand for the number of CNN filters in each residual connection, skip connection, and dilated convolutional block.

<table>
<thead>
<tr>
<th>Model</th>
<th>N stacks / layers</th>
<th>k filters</th>
<th># params</th>
<th>receptive field</th>
<th>target field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wave-U-Net</td>
<td>-</td>
<td>≈ 10.2M</td>
<td>9.21 s</td>
<td>1.02 s</td>
<td></td>
</tr>
<tr>
<td>DeepConvSep</td>
<td>1 stack / 10</td>
<td>512</td>
<td>128 ms</td>
<td>100 ms</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 stacks / 20</td>
<td>256</td>
<td>256 ms</td>
<td>100 ms</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 stacks / 30</td>
<td>128</td>
<td>384 ms</td>
<td>100 ms</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 stacks / 40</td>
<td>64</td>
<td>512 ms</td>
<td>100 ms</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 stacks / 50</td>
<td>32</td>
<td>639 ms</td>
<td>100 ms</td>
<td></td>
</tr>
</tbody>
</table>

Perceptual tests were conducted with 15 participants to get subjective feedback. Five songs were randomly chosen from 1’ to 1’10” to compose the perceptual test set. Participants were asked to “give an overall quality score, taking into consideration both: sound quality of the target source and interferences from other sources” for each of the estimated sources. The original mixture and the clean target source were presented as references. Participants provided a score between 1–5, with 1 being “very intrusive interferences from other sources and degraded audio”, and 3 being “noticeable interferences from other sources and not degraded audio”. Mean opinion score (MOS) is obtained by averaging the scores from all participants.

Table 2 shows the results of our experiments. Wide (but less deep) architectures fail at solving the task, only models having more than 3 stacks are able to perform competently. Two reasons may exist for that: wide models (having > 10M parameters) overfit the training set, and/or the small receptive field of wide models is not enough to solve the task. Further, we observe that the dissimilarity loss term $L_d$ does not help improving the results. Consequently, we choose the best...
training examples contain voice. Consequently, we choose the best performing model (4 stacks + $L_d$) for the perceptual test. Table 3 presents the results of the perceptual test, showing that participants preferred Wave-U-Net separations over Wavenet-based ones (t-test: p-value=0.049). This trend is consistent with BSS Eval scores, which denotes how powerful U-net architectures are for source separation [3, 20]. That said, the remarkable performance of the proposed Wavenet-based model also indicates the potential of end-to-end music source separation models in general. Informal listening also reveals that Wavenet-based models seem to better remove the accompanying sources. Although Wave-U-Net has difficulties in producing silences in parts having only one separation, its separations are smoother and have less artifacts — that’s why these separations are preferred by the listeners. Finally, end-to-end models trained only for singing voice separation achieve much better results than their counterparts trained for multi-instrument separation (compare Tables 4 and 5, against Tables 2 and 3).

5. Discussion

Throughout the years, end-to-end music source separation has been considered a hard research problem. Possibly because waveforms are variable and high-dimensional, the research community has focused on processing spectrograms instead of waveforms. However, with the recent advances of deep learning, music source separation starts to be possible in the waveform domain. As seen, although end-to-end music source separation methods have only started to be explored, the encouraging results we report denote the potential of this research direction — that might, e.g., allow to bypass the inherent phase problems associated with some spectrogram-based methods, or to move beyond the current mask-based filtering paradigm. To further show the viability of this research direction, we proposed a novel end-to-end source separation model based on Wavenet, that performs comparably to Wave-U-Net. However, these two state-of-the-art waveform-based models perform $\approx 1.5$dB (SDR) worse than the best spectrogram-based models that were published during the last SiSEC (Signal Separation Evaluation Campaign) [1]). Hence, although being possible and conceptually promising, end-to-end music source separation is still a challenging research topic. Finally, as an additional way to validate the direction we explored, it is worth mentioning that the speech source separation community is also starting to propose end-to-end methods with some degree of success [29, 30, 31].
7. References


