AN RNN MODEL FOR SINGLE CHANNEL SOURCE SEPARATION WITH ITERATIVE SUBTRACTION

Peter Li[1*], *Israel Malkin*[2], *Tian Wang*[3], *Kyunghyun Cho*[4], *Juan Bello*[1]

[1]New York University,[2]Butterfly Network, [3]eBay

ABSTRACT

Single channel source separation is the problem of recovering source components from a single channel mixture. It is a fundamental task in signal processing with many applications. In this paper, we propose a source separation model based on recurrent neural networks and a novel iterative subtraction architecture. We describe architectures and weight sharing methods for estimating sources via masks and spectrum directly. Our approach achieves a 5 dB - 7 dB SDR increase, 9.7 dB - 11.0 dB SIR increase, and 1.1 dB - 3.9 dB SAR increase over a NMF baseline in a closed speaker set evaluation. Further, we show that our proposed model is robust to additional noise and mixing conditions not seen during model training.

Index Terms— Source Separation, Deep Learning

1. INTRODUCTION

Single channel source separation is the problem of recovering source components from a single channel mixture. It is a fundamental task in signal processing with many applications including robust ASR, speaker identification, and hearing prothesis. Without prior information,however, this is an under determined problem with an infinite number of solutions. Data driven approaches attempt to learn a separation model using data as prior information. For example, early data driven approaches used non-negative matrix factorization (NMF) and probabilistic latent semantic indexing (PLSI) to factorize and separate a time-frequency representation of the mixture using learned source basis vectors [1][2]. Fundamentally, these rely on learning linear transformations to perform separation. In recent years, deep learning approaches have been proposed to leverage the ability of deep learning systems to learn flexible, non-liner models. The development has generally fallen into two groups. Works such as [3] and [4] treat source separation as a regression problem where the model produces estimates of the individual sources. Generally, these models take the mixture and output a pre-determined number of sources. Other methods treat separation as a clustering problem where the goal is to cluster time-frequency bins belonging to the same source together [5][6]. These methods typically learn an embedding for each each time-frequency bin and cluster this learned representation. These models are highly flexible and can be use to produce multiple sources (by requiring more clusters). See [7] for an in-depth survey of deep learning methods for speech source separation.

In this paper, we propose a source separation model based on recurrent neural networks and a novel iterative subtraction architecture. This model allows us to train a regression based deep learning system that is not limited to producing a predetermined number of outputs. We show that our model outperforms a NMF baseline and generalizes on mixtures with different mixture levels and SNR.

2. PROPOSED METHODS

For this work, we focus on the task of separating additive mixtures of N sources:.

$$y(t) = \sum_{i=1}^{N} x_i(t) \tag{1}$$

Given the mixed signal $y(t)$, we wish to estimate the individual sources, $x_i(t)$. This work focuses on mixtures of speech signals, but the methods discussed may be extended to other signal types such as music or environmental sounds.

2.1. Model

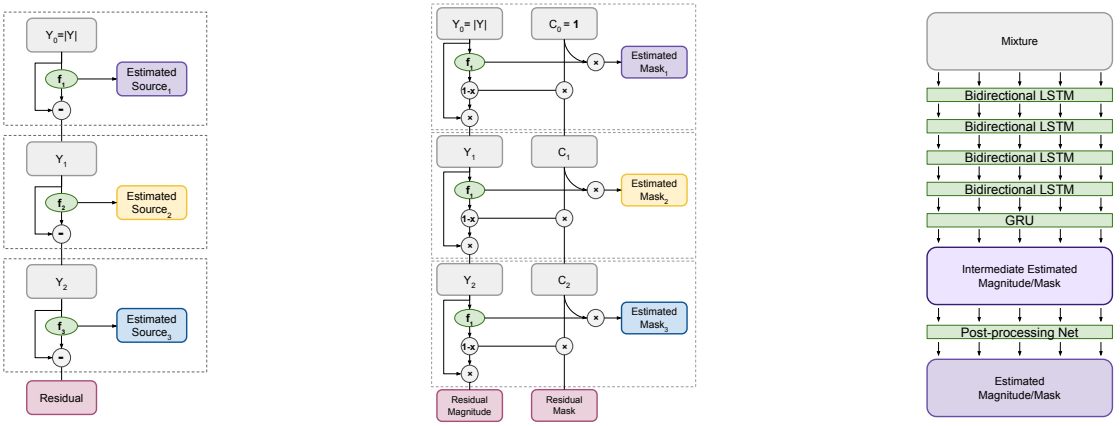
For separation, we consider the time-frequency representation of Equation 1. Let $\mathbf{Y}, \mathbf{X}_i \in \mathbb{C}^{F \times T}$ be the short-time Fourier transform (STFT) of $y(t)$ and $x_i(t)$ respectively. Assuming additivity of α-spectrograms[8], we can decompose the complex modulus of $\mathbf{Y}$ as:

$$|\mathbf{Y}(t,f)|^{\alpha} \approx \sum_{i=1}^{N} |\mathbf{X}_i(t,f)|^{\alpha} \tag{2}$$

Empirical results in [8] show that $\alpha \approx 1$ best fits real-world data. In this work, we assume additivity with $\alpha = 1$.

We propose a method to separate the mixture via *subtraction*. We give an overview of the idea below and give detailed

*This work was partially supported by National Science Foundation award 1544753.



(a) Magnitude Model (b) Mask Model (c) Separator RNN

Fig. 1: Model Architecture. (a) and (b) shows an example for a three speaker mix. (c) shows the architecture of the separator.

descriptions in the following three sections. Our model processes data sequentially on two levels:

1. Given a mixture spectrogram, we use a separator, f_i, to separate out *one* source spectrogram. We use a recurrent neural network (RNN) to process the spectrogram sequentially. See section 2.4 for a detailed description.

2. To estimate multiple sources, we process the mixture sequentially. Each separator takes in a mixture less the estimated source from the preceding separator. See section 2.2 and section 2.3 for two proposed models.

For any given model, there are two ways to specify the separator weights. In one case, we have different separators for each layer. We term this the speaker dependent (SD) model since the separator is tied to its layer and the data it receives. The benefit of this architecture, however, comes when we force the separators to share weights. We term this the speaker independent (SI) model since the separator is now independent of its place in the architecture. The iterative subtraction allows us to train a general purpose separator that is more independent to characteristics of the mixture e.g. speaker identity, gender, and number of sources.

2.2. Magnitude Spectrum Model

In our first model, the separator produces estimates of the source spectrum directly.The initial input to the model is the full mixture spectrum $Y_0 = |Y|$. The i^{th} source is then computed using the following recurrence:

$$\hat{X}_i = f_i(Y_i|\theta_i) \tag{3}$$

$$Y_i = Y_{i-1} - \hat{X}_{i-1} \tag{4}$$

Additionally, ϵ is the residual after the last estimated source has been subtracted, $\epsilon = Y_N$. This captures any part of the original mixture spectrogram that was not output as part of the source. See Figure 1a for an example for a 3 speaker mix.

For model training, we optimize the separator parameters jointly to minimize the reconstruction loss and the norm of the residual. λ controls the relative importance of the reconstruction and residual loss.

$$\underset{\theta_1,\ldots,\theta_N}{\text{minimize}} \quad \sum_i^N \|X_i - \hat{X}_i(\theta_i)\|_F^2 + \lambda\|\epsilon\|_F^2 \tag{5}$$

To reconstruct the time domain signal, we take the inverse short-time Fourier transform (ISTFT) using the estimated magnitude spectrograms $\hat{X}_i$, and the mixture phase.

2.3. Soft-mask Model

In addition to estimating source magnitude spectrograms directly, we can also recover $|X_i|$ by multiplying $|Y|$ by a phase adjusted soft mask M_i defined as:

$$M_i := \frac{|X_i|}{\sum_{j=1}^N |X_j|} cos(\phi) \tag{6}$$

.

where ϕ is the difference between the mixture and source phase. This adjustment accounts for the phase error when reconstructing using the mixture phase and has been found to lead to better SDR [9].

For masks, the sequential processing is slightly more complex as we also need to maintain memory of previous masks. As in the magnitude model, the initial input is the full mixture spectrum $Y_0 = |Y|$. Additionally, we introduce a memory component that is initialized as the identity mask, $C_0 = \mathbb{1}$. The i^{th} mask is computed with the following recurrence:

$$\tilde{M}_i = f_i(Y_i|\theta_i) \tag{7}$$

$$\hat{M}_i = \tilde{M}_i \odot C_i \tag{8}$$

$$Y_i = Y_{i-1} \odot (1 - \tilde{M}_{i-1}) \tag{9}$$

$$C_i = C_{i-1} \odot \tilde{M}_{i-1} \tag{10}$$

Here, $\tilde{M}_i$ is a *local* mask in the sense that it is used to mask its input, Y_i. In later layers, separators do not have access to sources that were already subtracted so a separator cannot produce a global mask on $|Y|$. The memory, C, remedies this by serving as a running mask of what has been subtracted.Intuitively, C_i is the mask on $|Y|$ to produce Y_i. A global mask can be recovered by $C_{i-1} \odot \tilde{M}_{i-1}$

The model is optimized in an analogous way to the magnitude model, where $\epsilon = C_n$ is the residual mask.

$$\underset{\theta_1,\ldots,\theta_N}{\text{minimize}} \quad \sum_i^N \|M_i - \hat{M}_i(\theta_i)\|_F^2 + \lambda\|\epsilon\|_F^2 \tag{11}$$

To reconstruct the time domain signal, we take the inverse short-time Fourier transform (ISTFT) using the estimated magnitude spectrograms $\hat{X}_i = |\mathbf{Y}| \odot \hat{\mathbf{M}}_i$, and the mixture phase.

2.4. Separator RNN

We consider spectrograms and masks as sequences of F-dimensional vectors e.g., $Y = (y_1, ..., y_T)$. The source separation problem can then be viewed as mapping a variable length sequence (mixture) into another sequence of the same length (source/mask). This is a natural task for recurrent neural networks (RNN) as they can handle variable length input and they can flexibly model temporal patterns in the data.

For our separator,we use four layers of bidirectional RNN with long short-term memory (LSTM) [10] units followed by a single RNN with gated recurrent units (GRU) [11] that outputs vectors of dimension F.

During our experiments, we found that spectrograms/masks output from the separator RNN could be further improved ($\sim$0.8 dB increase in SDR) with a post-processing network. The post-processing network follows the architecture first proposed in [12] and used for similar post-processing in [13]. See Table 1 for detailed description of the architectures and model hyper-parameters.

3. EXPERIMENTS

In our experiments we examine the performance of our model in three settings:new utterances, different mixture SNR, and additional noise. As a proof on concept, our experiments are run on a closed speaker set, i.e. both training and test data are from the same set of speakers.

3.1. Data and Evaluation

We use speech samples from the TSP Speech Database[14]. We choose six speakers: three male {`MA, MB, MC`} and three female {`FA, FB, FC`}. For each speaker, we use a 70/10/20 split on utterances for train, validation, and test splits. All audio signals are downsampled to 16 kHz. We use a 1024-point STFT with 75% overlap and a Hann window. Mixtures are created by mixing one male and one female utterance at 0 dB. For evaluation we use three common metrics proposed in [15]: source to interference ratio (SIR), source to artifacts ratio (SAR) and source to distortion ratio (SDR). These metrics measure the ratio in power of the clean source to various components of the estimated source. Higher values tend to represent higher separation quality. See [15] and [3] for a more in depth discussion.

3.2. New Utterances

We fist evaluate the performance of our models on mixtures of the test utterances mixed at 0 dB. We use a standard supervised NMF with KL divergence as a baseline. The separation results are shown in Table 2.

	SDR	SIR	SAR
Supervised NMF	5.20	7.35	10.13
Deep Attractor Network	12.75	19.00	14.10
Magnitude Spectrogram Model			
Speaker Dependent	10.20	17.80	11.21
Speaker Independent	10.75	18.35	11.73
Soft-mask Model			
Speaker Dependent	12.35	17.75	14.03
Speaker Independent	11.92	17.01	13.78

Table 2: Speech separation results. The mask model out performs the magnitude model. Additionally, speaker dependent and speaker independent models show similar performance.

We see that the mask models outperforms the magnitude models and both groups outperform the NMF baseline. The higher mask predictions may be because their predictions are based on masked values of the mixture spectrogram. This is much stronger prior information than the the spectrogram model when needs to predict values from $\mathbb{R}_+$. Additionally, we observe that the SI models performs on par with the SD model in both the spectrogram and mask models despite only having one separator. This means the single SI separator was able to separate on both male and female voices. We also include the performance of a pre-trained Deep Attractor Network (DANet). The comparison is not entirely equal, since DANet was trained on different data. We report the result to represent the current state of the art.

Model Component	Architecture and Hyper-parameters
Separator RNN	4 layers of bidirectional LSTM, hidden units: 500 (speaker dependent), 1000 (speaker independent) GRU: hidden units: 513
Post-processing RNN	1-D Convolutional filter bank: kernel width = 1-16, 128 output channels per width, stride = 1, ReLU Max Pool: stride = 1, width = 2 1-D Convolutional filter bank: width=3, 128 output channels, stride = 1 ReLU 1-D Convolutional filter bank: width=3, 128 output channels, stride = 1 Residual connection to input Highway Net: 4 layers of fully connected layers, 128 units, ReLU Bidirectional GRU: 128 hidden units, ReLU (magnitude), sigmoid (mask)

Table 1: Network Architectures and hyper-parameters. Batch normalization used between all convolution layers in the post-net.

3.3. Different SNR

Using the speaker independent mask model from the preceding section, we next evaluate separation performance on test mixtures where the sources are combined at different mixture SNR (the difference in power between x_1 and x_2). The new test mixture is defined as: $y(t) = \alpha x_1(t) + x_2(t)$ where α is the gain that controls the relative loudness.

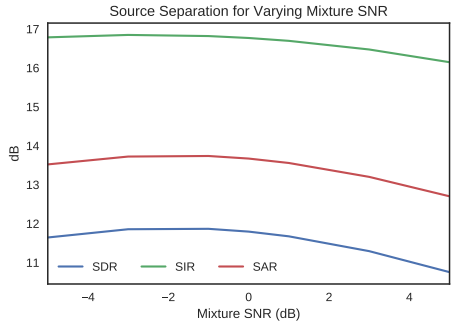


Fig. 2: Speech separation results for different mixture SNR. Separation quality is robust to different SNR.

From Figure 2, we see that the evaluation metrics vary slightly through a range of mixture SNR from -5 dB to 5 dB. This shows that our separator is robust to a range of mixing even though the model was only trained on mixes at 0 dB. This is consistent with the finding in the speech enhancement literature where supervised speech enhancement algorithms are not overly sensitive to differences in SNR. Wang posits that this may be due to loudness differences that occur naturally within an audio sample[7].

3.4. Noise

Next, we evaluate separation performance in the presence of noise. Noisy mixtures are defined as: $y(t) = x_1(t) + x_2(t) + e(t)$ where $e \sim N(0, \sigma)$ is Gaussian white noise and σ controls the signal to noise ratio (SNR). Given the noisy mixture, we evaluate separation performance against the clean sources $x_1(t)$ and $x_2(t)$.

In Figure 3 we see that SDR and SAR decrease as SNR decreases. Intuitively this makes sense since separation will be harder as noise levels rise. Interestingly, we do not see a complete collapse in performance, even though the models were never trained on noisy mixes.

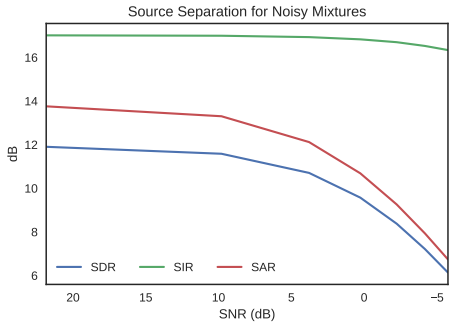


Fig. 3: Speech separation results on noise mixtures.Up to 10 dB SNR, separation quality is relatively robust to added noise. Beyond 10 dB separation performance decreases sub linearly. However, separator performance does not collapse even though it was trained on clean mixtures.

4. CONCLUSION

We propose a model for single channel source separation using recurrent neural networks and an iterative subtraction architecture. We demonstrate the training of both speaker dependent and independent models for mask and magnitude prediction. We evaluate these models under a number of conditions on a closed speaker set and show that they outperform a NMF baseline and slightly lower than a pre-trained DANet.

Preliminary experiments on open speaker TSP data show that our model has lower performance (8.1 dB SDR for a SI model). In future work, we would like to explore model behavior on larger, open speaker data. Additionally, adversarial training or different loss functions e.g. PIT loss may yield better separation.

5. REFERENCES

[1] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[2] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Independent Component Analysis and Signal Separation*, pp. 414–421, 2007.

[3] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2136–2147, 2015.

[4] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Joint separation and denoising of noisy multi-talker speech using recurrent neural networks and permutation invariant training," *arXiv preprint arXiv:1708.09588*, 2017.

[5] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[6] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 246–250.

[7] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: an overview," *arXiv preprint arXiv:1708.07524*, 2017.

[8] Antoine Liutkus and Roland Badeau, "Generalized wiener filtering with fractional power spectrograms," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 266–270.

[9] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 708–712.

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

[11] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[12] Jason Lee, Kyunghyun Cho, and Thomas Hofmann, "Fully character-level neural machine translation without explicit segmentation," *arXiv preprint arXiv:1610.03017*, 2016.

[13] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech syn," *arXiv preprint arXiv:1703.10135*, 2017.

[14] Peter Kabal, "Tsp speech database," .

[15] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.