Monaural Source Separation using Deep Recurrent Neural Networks

Anand Parwal *

Abstract— **Monaural source separation is the challenging task of separating signals from different sources when only one mixed channel information is available. Even though people can intuitively distinguish a conversational thread amid a cacophony of babble in a crowded room, it's not a trivial task algorithmically. In this project, the capability of deep recurrent neural networks (RNN) for singing voice separation from monaural recordings in a supervised manner is studied. Different architectures for Deep recurrent neural networks and different objectives are explored.**

I. Introduction

In natural conversation a speech signal is typically perceived against a background of other sounds (background noise, music, other speech). The human auditory system processes the acoustic mixture reaching the ears to enable constituent sounds to be heard and recognized as distinct entities, even if these sounds overlap in both spectral and temporal regions with the target speech. The flexibility and robustness of human speech perception is demonstrated by the range of situations in which spoken communication is possible in the presence of competing sound sources [1]. Researchers in signal processing and many other related fields have strived for the realization of this human ability in machines; however, it is still a topic of research.

Monaural source separation is important for several real world applications. For example, the accuracy of automatic speech recognition (ASR) can be improved by separating noise from speech signals [2]. The accuracy of chord recognition and pitch estimation can be improved by separating singing voice from music [3]. It can be used to recognize speech from different voices in a single audio clip, known as the cocktail party problem. However, current state-of-the-art results are still far behind human capability. In this project, I focused on singing voice separation from monaural recordings as a test case for monaural source separation.

The organization of this paper is as follows: Section 2 discusses the relation to previous work. Section 3 introduces the proposed methods, including the deep recurrent neural networks, joint optimization of deep learning models and a soft time-frequency masking function, and different training objectives. Section 4 presents the experimental setting and results using the MIR-1K dataset [12]. The paper is concluded in Section 5.

II. Literature Review

Various sophisticated methods have been proposed over the past few decades in research areas such as computational auditory scene analysis (CASA) [4] and independent component analysis (ICA) [5–7]. CASA separation techniques are mostly based on splitting mixtures observed as a single stream into different auditory streams by building an active scene analysis system for acoustic events that occur simultaneously in the same spectro-temporal regions. The acoustic events are distinguished according to rules inspired intuitively or empirically from the known characteristics of the sources.

Nonnegative matrix factor deconvolution (NMF) [8, 9] can extract an inherent spectro-temporal structure of a sound source. As a result of NMF, a dictionary of monotonic trajectories is learned from a course of sound source power spectral densities, and by classifying the dictionary items into a desired number of elements, the original source signals can be recovered. Although NMF is successfully applied to several monaural source separation problems such as polyphonic music transcription, it is hard to obtain a reliable dictionary for a complex sources such as speech signals

The recent success of deep neural networks for classification problems has naturally inspired their use in class-based segmentation problems and therefore highly specialized algorithms such as CASA requiring advanced knowledge of audio signal processing have seen a decrease in use.

One of the major problem for deep learning algorithms is that many of the neural networks developed assume the number of speakers present in the mixture, which is necessary considering the number of output nodes in a neural network is fixed and may not match the total number of speakers in the audio. Even though spectral clustering techniques [10] can accommodate unknown number of speakers, they are limited in performance as they are dependent on specially designed features and don't utilize the capabilities of deep learning. One proposed way to solve this problem is using long short-term memory (LSTM) layers in a deep clustering approach [11] to learn feature transformations known as embeddings, which can then be used for clustering. This approach can represent the various speakers implicitly using the fixed-dimensional output of the network.

*Anand Parwal is with the Robotics Department, Worcester Polytechnic Institute, Worcester, MA 80305 USA (e-mail: aparwal@ wpi.edu).

Project explores different deep recurrent neural network architectures along with the joint optimization of the network with a soft masking function.

III. Problem Description

In order to formulate the problem, assume that the observed signal y^t is the summation of N independent source signals

$$y^t = \lambda_1 x_1^t + \lambda_2 x_2^t + \ldots + \lambda_N x_N^t, \qquad (1)$$

where x_i^t is the t-th observation of the i-th source, and λ_i is the gain of each source, which is fixed over time. Note that superscripts indicate sample indices of time-varying signals and subscripts identify sources. The gain constants are affected by several factors, such as powers, locations, directions and many other characteristics of the source generators as well as sensitivities of the sensors.

The goal is to recover all x_i^t given only a single sensor input y^t. The problem is too ill-conditioned to be mathematically tractable since the number of unknowns is NT given only T observations

A. Dataset

The proposed architecture is evaluated using the MIR-1K dataset [12]. A thousand song clips are encoded with a sample rate of 16 KHz, with durations from 4 to 13 seconds. The clips were extracted from 110 Chinese karaoke songs performed by both male and female amateurs. There are manual annotations of the pitch contours, lyrics, indices and types for unvoiced frames, and the indices of the vocal and non-vocal frames. Only the singing voice and background music are used in this set-up.

Data is generated in 4 second chunks and training data is augmented by combining vocals of a song with music from another. The network is trained using clips from 50 singers, validated on 20 singers and tested on remaining 40 singers' clips

B. Evaluation

Following the evaluation framework in [12], three sets of mixtures were created using the 1000 clips of the MIR-1K dataset. For each clip, the singing voice and the music accompaniment were mixed at -5, 0, and 5 dB SNRs, respectively. Zero indicates that the singing voice and the music are at the same energy levels, negative values indicate the energy of the music accompaniment is larger than the singing voice, and so on.

For source separation evaluation, in addition to evaluating the Global Normalized Source to Distortion Ratio (GNSDR) as [12], performance is also evaluated in terms of Source to Interference Ratio (SIR), Source to Artifacts Ratio (SAR), and Source to Distortion Ratio (SDR) by BSS-EVAL metrics [13]. The Normalized SDR (NSDR) is defined as

$$NSDR(\hat{v}, v, x) = SDR(\hat{v}, v) - SDR(x, v), \qquad (2)$$

where $\hat{v}$ is the resynthesized singing voice, v is the original clean singing voice, and x is the mixture. NSDR is for estimating the improvement of the SDR between the preprocessed mixture x and the separated singing voice $\hat{v}$. The GNSDR is calculated by taking the mean of the NSDRs over all mixtures of each set, weighted by their length.

$$GNSDR(\hat{}v, v, x) = \frac{\sum_{n=1}^{N} w_n NSDR(\hat{v}_n, v_n, x_n)}{\sum_{n=1}^{N} w_n}, \qquad (3)$$

where n is the index of a song and N is the total number of the songs, and w_n is the length of the nth song. Higher values of SDR, SAR, SIR, and GNSDR represent better separation quality.

IV. Proposed Method

A. Recurrent Neural Network

To capture the contextual information among audio signals, one way is to concatenate neighboring features together as input features to the deep neural network. However, the number of parameters increases rapidly according to the input dimension. Hence, the size of the concatenating window is limited. A recurrent neural network (RNN) can be considered as a DNN with indefinitely many layers, which introduce the memory from previous time steps. The potential weakness for RNNs is that RNNs lack hierarchical processing of the input at the current time step. To further provide the hierarchical information through multiple time scales, deep recurrent neural networks (DRNNs) are explored.

B. Time Frequency Masking

Given the input features, x_t from the mixture, we obtain the output predictions $\hat{y}_{1_t}$ and $\hat{y}_{2_t}$ through the network. The soft time-frequency mask m_t is defined as follows:

$$m_t(f) = \frac{|\hat{y}_{1_t}(f)|}{|\hat{y}_{1_t}(f)| + |\hat{y}_{2_t}(f)|}, \qquad (4)$$

where f ∈ {1, ... , F} represents different frequencies.

Once a time-frequency mask m_t is computed, it is applied to the magnitude spectra z_t of the mixture signals to obtain the estimated separation spectra$\hat{s}_{1_t}$ and $\hat{s}_{2_t}$, which correspond to sources 1 and 2, as follows:

$$\hat{s}_{1_t}(f) = m_t(f)\, z_t(f) \qquad (5)$$

$$\hat{s}_{2_t}(f) = (1 - m_t(f))\, z_t(f)$$

where f ∈ {1, . . . , F} represents different frequencies.

The time-frequency masking function can be viewed as a layer in the neural network as well. Instead of training the network and applying the time-frequency masking to the results separately, we can jointly train the deep learning models with the time-frequency masking functions. We add an extra layer to the original output of the neural network as follows:

$$\bar{y}_{1_t} = \frac{|\hat{y}_{1_t}|}{|\hat{y}_{1_t}| + |\hat{y}_{2_t}|} \circ z_t \qquad (6)$$

$$\bar{y}_{2_t} = \frac{|\hat{y}_{2_t}|}{|\hat{y}_{1_t}| + |\hat{y}_{2_t}|} \circ \; z_t$$

where the operator ° is the element-wise multiplication (Hadamard product). In this way, we can integrate the constraints to the network and optimize the network with the masking function jointly. Note that although this extra layer is a deterministic layer, the network weights are optimized for the error metric between and among $\bar{y}_{1_t}$, $\bar{y}_{2_t}$ and y_{1_t} , y_{2_t} , using back-propagation. To further smooth the predictions, we can apply masking functions to $\bar{y}_{1_t}$ and $\bar{y}_{2_t}$, as in Eqs. (4) and (5), to get the estimated separation spectra $\bar{s}_{1_t}$ and $\bar{s}_{2_t}$. The time domain signals are reconstructed based on the inverse short time Fourier transform (ISTFT) of the estimated magnitude spectra along with the original mixture phase spectra

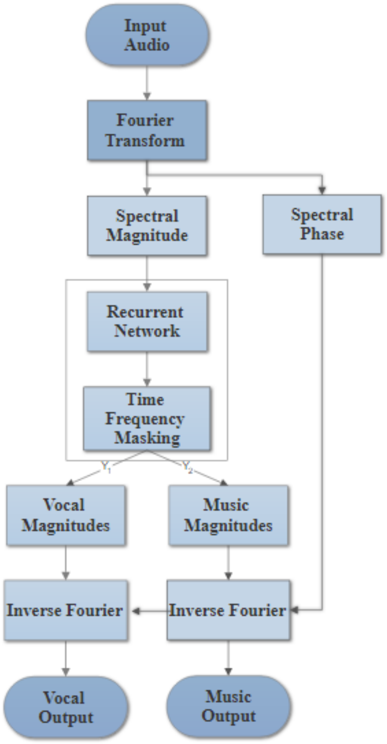

Figure 2. Proposed framework

C. Architecture

At time t, the training input, x_t, of the network is the concatenation of features from a mixture within a window. Magnitude spectra is used as features. The output targets, y_{1_t} and y_{2_t}, and output predictions, $\bar{y}_{1_t}$ and $\bar{y}_{2_t}$, of the network are the magnitude spectra of different sources. Since the problem is to separate one of the sources from a mixture, instead of learning one of the sources as the target, the framework from [14] is adapted to model all different sources simultaneously. Figure 1 shows an example of the architecture. Moreover, it is useful to further smooth the source separation results with a time-frequency masking technique, for example, binary time-frequency masking or soft time frequency masking [14]. The time-frequency masking function enforces the constraint that the sum of the prediction results is equal to the original mixture. For manageability all hidden layers had 256 hidden units.

V. Experiments and Results

In the separation process, the spectrogram of each mixture is computed using, a window size of 1024, short time Fourier transform (STFT) with a hop size of 256 (at Fs=8,000). Using log-mel filterbank features provided worse performance. Many experiments were performed by changing the number of RNN layers as 1, 2 and 3, using loss as MSE, the mean squared error, and KL (the generalized Kullback-Leibler divergence criterion) and using a discriminative training objective and changing the input context size. Unless stated, training was done on 1000 epochs with batch size of 1. Optimizer adam was found to converge satisfactorily within 1000 epochs, hence it was used in all expriments.

A. Dataset

The proposed architecture is evaluated using the MIR-1K dataset [12]. A thousand song clips are encoded with a sample rate of 16 KHz, with durations from 4 to 13 seconds. The clips were extracted from 110 Chinese karaoke songs performed by both male and female amateurs. There are manual annotations of the pitch contours, lyrics, indices and types for unvoiced frames, and the indices of the vocal and non-vocal frames. Only the singing voice and background music are used in this set-up.

Data is generated in 4 second chunks and training data is augmented by combining vocals of a song with music from another. The network is trained using clips from 50 singers, validated on 20 singers and tested on remaining 40 singers' clips.

B. Results

The suppression of noise is reflected in SIR. The artifacts introduced by the denoising process are reflected in SAR. The overall performance is reflected in SDR.

First, the effect of input size is compared. Using only the current frame, one previous frame and two previous frames is compared with one RNN layer and MSE loss after time frequency masking. Table 1 shows that using one previous frame worked best and therefore in later experiments, input size of two is used.

TABLE I. Results for input sizes

Model (Input Size)	GNSDR	GSIR	GSAR
1	6.23	9.34	9.83
2	6.55	9.45	9.89
3	6.46	9.45	9.91

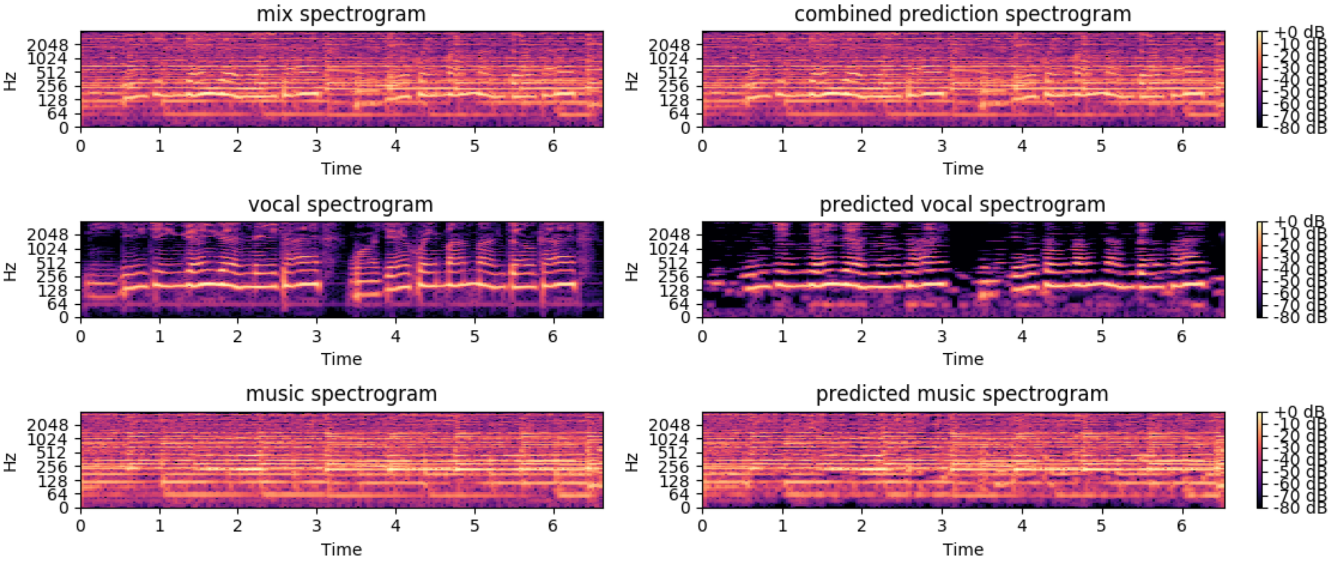

Figure 2. Comparision of magnitude of spectogram of vocal and music audio for original tracks (left) and predicted tracks (right) for file davidson_3_11.wav in MIR-1K dataset using MSE loss and 3 layer RNN with joint time frequency optimization.

Next, performance with and without data augmentation is compared. For augmentation, every 4 second clip was combined with 10 other music tracks, increasing the training set ten times.

As Table II shows, augmentation provided significant improvement over original data and was used in all later experiments

TABLE II. RESULTS FOR DATA AUGMENTATION

Input	GNSDR	GSIR	GSAR
Not Augmented	6.55	9.45	9.89
Augmented	7.13	10.03	10.64

Next, performance for different targets for loss was compared. As seen in Table III, when both the sources are used for target, loss can be computed after or before the soft time frequency mask was applied. Using only one source as a target gave considerably worse performance while computing loss after the mask was still the best performing model.

TABLE III. RESULTS FOR TARGETS FOR LOSS COMPUTATION

Training loss computed on (output)	GNSDR	GSIR	GSAR
Vocal	5.63	8.12	9.01
Music	5.32	8.32	8.88
Both, before mak	6.89	9.71	9.02
Both, after mask	7.13	10.03	10.64

Next, performance for different architectures of the model and objective functions was explored, with number of RNN layers ranging from 1 to 3 and simply DNN. For all the models, use of KL divergence and MSE was compared.

TABLE IV. RESULTS FOR DIFFERENT MODEL ARCHITECTURE AND LOSSES

Model	Loss	GNSDR	GSIR	GSAR
DNN	MSE	6.18	8.92	9.11
1 RNN layers	MSE	7.13	10.03	10.64
2 RNN layers	MSE	7.18	10.71	10.92
3 RNN layers	MSE	7.22	10.53	11.04
DNN	KL	6.53	8.50	9.48
1 RNN layers	KL	7.13	10.01	10.78
2 RNN layers	KL	7.20	10.28	11.19
3 RNN layers	KL	7.25	10.46	11.15

As Table IV clearly shows KL divergence criterion always performed marginally better than MSE. Increasing the number of RNN layers to 3 gave better SDR values

Finally, comparing this result with the baseline NMF results [8], there is an improvement of 2.4 dB GNSDR.

VI. CONCLUSION

Different deep learning models for monaural source separation were studied and compared. Recurrent neural networks and joint time frequency mask optimization was used to separate voice and music from mixed audio and an improvement over the baseline was achieved.

REFERENCES

[1] A. S. Bregman. “Auditory scene analysis: The perceptual organization of sound”. MIT press; 1994.

[2] A. L. Maas, Q. V Le, T. M O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng. "Recurrent neural networks for noise reduction in robust ASR," INTERSPEECH, 2012

[3] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 57–60, 2012

[4] G. J. Brown and M. Cooke, "Computational auditory scene analysis," Computer Speech and Language, vol. 8, no. 4, pp. 297–336, 1994.

[5] P. Comon, "Independent component analysis, A new concept?" Signal Processing, vol. 36, pp. 287–314, 1994.

[6] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," Neural Computation, vol. 7, no. 6, pp. 1004–1034, 1995.

[7] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," IEEE Trans. on S.P., vol. 45, no. 2, pp. 424–444, 1996

[8] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in Proc. ICA 2004, vol. 3195, pp. 494–501, Sept. 2004.

[9] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in Proc. ICA 2006, Apr. 2006

[10] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," JMLR, vol. 7, pp. 1963–2001, 2006

[11] J. Hershey, Z. Chen, J. Roux, and S. Watanabe. "Deep clustering: Discriminative embeddings for segmentation and separation." In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016

[12] C. L. Hsu and J.S.R Jang. "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset". IEEE Transactions on Audio, Speech, and Language Processing, 18(2), pp.310-319, 2010.

[13] R. Gribonval, L. Benaroya, E. Vincent, C. Fvotte, "Proposals for performance measurement in source separation", Proc. Int. Symp. ICA BSS, pp. 763-768, 2003-Apr

[14] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Deep learning for monaural speech separation. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.