Implementation Singing-Voice Separation

E4040.2016Fall.HINT.report
Chang Pan cp2923, Kejia Shi ks3403, Shengyang Zhang sz2624
Columbia University

We first tried studying another paper *A Neural Algorithm of Artistic Style.* We hit rock when implementing VGG19 model without using existing packages. This has already wasted us more than half the time. We moved on studying this paper with the rest of time.

Abstract

We implement the method of using Deep Recurrent Neural Network (DRNN) to separate singing voice from monaural recordings and achieve a GNSDR score of 6.08 and 6.13. In addition, we propose several improvements to the original algorithm, including using curriculum learning, optimizing parameters with the Itakura-Saito distance. We train and test our model on the MIR-1K dataset.

1. Introduction

Given a monophonic track mixed with singing voice and background sound, it can be essential to separate them in real life to extract useful components. For instance, measuring vocal pitch needs a clear separation of vocal from background music. Better recording performance would better isolate noise from true source.

The recent spring-up deep learning methodology showcases a new perspective from traditional linear transformation in solving this problem, which introduces extra constraints. Bayesian method is a solution to avoid the constraint issue, but it varies in situations and could be slow. (Yang et al., 2014)

Numerous papers using neural networks have come up these years. One way of using deep learning framework to finish our task is to use conventional regression-based networks to infer the source signals directly, using time-frequency (T-F) masks. Huang et al. (2014) proposes such a framework using DRNN optimized by a soft mask function on this task. Other ways such as deep clustering could be an useful approach to solve the problem as well. (Hershey, 2016) We explore potential adaptations based on Huang et al. (2014).

We deliver our project results as follows: Section 2 discusses the understanding of the original paper. Section 3 introduces our proposed methods, including the objectives and technical challenges, the break-down of the problem. Section 4 presents the detailed implementation of our proposal, including model parameters. Section 5 gives the results of our model in comparison with the original paper. We conclude the paper in Section 6.

2. Summary of the Original Paper

2.1 Methodology of the Original Paper

Huang et al. (2014) implements a DRNN structure in expectation of capturing the audio signals' neighboring features together as hierarchical inputs of the deep neural network, which is more informative. The general framework from the original paper is depicted in Figure 1-1 and 1-2. After a short-time Fourier transformation (STFT), the magnitude spectra are sent directly to the neural network built.

The soft time-frequency masking function is jointly trained with the networks to further smooth the separation results, by limiting the predicted component sum the same as the original mixture. It is realized by adding an extra layer to the original output.

Optimization of the neural network parameters is based on minimizing the mean square errors and the generalized KL-Divergence.

Lastly, the estimated magnitude spectra are sent to an inverse short time Fourier transform (ISTFT). Combined with the original mixture phase spectra, the time domain signals are recovered.

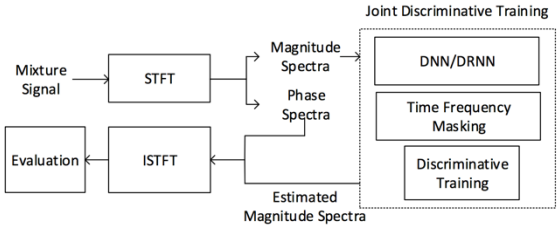

Figure 1-1 Huang et al. (2014) Framework

2.2 Key Results of the Original Paper

Huang et al. (2014) experiments with both Deep Neural Network and Deep Recurrent Neural Network parameters. In their findings, DNN with context window size 3, circular shift step size 10,000, two sources output with joint mask achieves a GNSDR score of 6.93.

In the comparison of different architectures, 2-layer DRNN with discriminate objective function, which

penalizes two sources and their cross prediction distances, achieves the best GNSDR score of 7.45.*

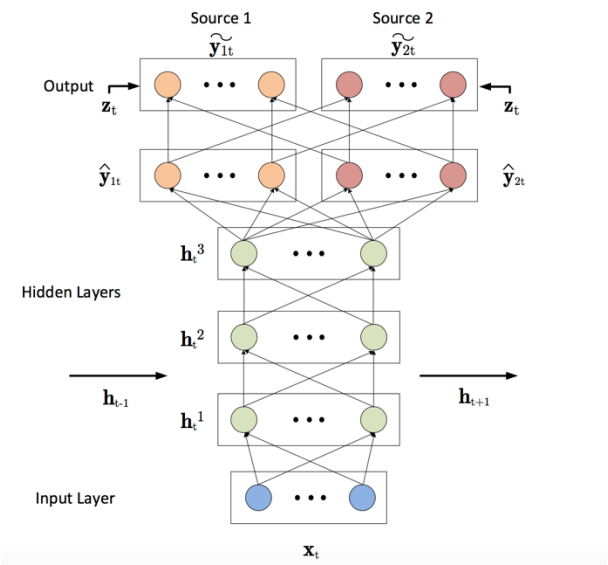

Figure 1-2 Huang et al. (2014) Neural Network

3. Methodology

3.1. Objectives and Technical Challenges

Our main objectives of this project are as follows: using what we have learned to train a functional deep recurrent neural network, specifically to the singing-voice separation framework; consolidating understanding of the signal transformation process, how deep learning better solves traditional problems in reducing redundant constraints, saving information and raising separation accuracy. More importantly, raising our own thinking of potentials for improvement.

We planned to build up the original model with two possible adaptations, using the Itakura–Saito distance as our optimization objectives and curriculum training to improve the training quality. With time constraint, we only successfully applied one as we only described the other promising direction.

The challenging part of the project lies in: building a functional neural network based on what we have learned – making sure the input, output dimensions are right; seeking better measurement to optimize; thoughtful consideration of improving the total training power.

* GNSDR: weighted Source to Distortion Ratio (SDR) score by lengths of music slices, which could be an indicator of the overall separation performance.

3.2. Problem Formulation and Design

Huang et al. (2014) uses mean squared error and KL-divergence as their optimization objectives. We adopt the Itakura–Saito distance here, which has many properties which are powerful and desirable. It has been shown that if the Itakura-Saito distance measure between two speech signals is less than 0.5, the difference in the mean opinion score would be less than 1.6. If the value is below 0.1, the original and the approximation would be perceived nearly identically by human ears. (Benesty, 2007)

While not having enough time putting into practice, we investigated in the theoretical possibility of adopting the curriculum learning method first raised by Bengio et al. (2009) in addition to the framework of Huang et al. (2014). This training method include two different segment lengths, 100 and 400. The shorter segments are successfully used for solving singing-voice separation tasks by pretraining and further training the longer ones, so as to improve the overall performance. (Isik, 2016) It should gain better performance when embedded to our training framework.

4. Implementation

4.1. Deep Learning Network

The Itakura–Saito distance is a measure of the difference between an original spectrum $P(\omega)$ and the approximation $Q(\omega)$ of that spectrum. It was proposed by Fumitada Itakura and Shuzo Saito in the 1960s. The mathematical formula is as follows:

$$\text{Itakura} - \text{Saito Distance} = \sum_i \log\frac{|P(\omega)|}{|Q(\omega)|} + \frac{|P(\omega)|}{|Q(\omega)|} - 1$$

The original model is proposed with 3 hidden layers with 1000 hidden units. We consider 1000 hidden units an excessive amount, which unnecessarily slows down the speed. In our experimentation, 150 units are enough to produce a reasonable result.

4.2. Software Design

We implement Huang et al. (2014) on the same MIR-1K dataset. The original dataset has 1,000 Chinese karaoke songs with vocals sung by 19 different people. For better comparison, we split our training and test set in the same way as the original paper, specifically 175 training samples and 825 testing samples. In addition, we pick 25 representative samples from the training set as our validation set, which downsizes the total training samples

to 150. We then augment the 175 samples for analysis first, and then put them to the algorithm.

We add a tiny noise to the magnitude value when performing normalization, in order to avoid having zero denominators when optimizing with divergence measurements.

The STFT and ISTFT are performed by using the python package `LibROSA`. We use the same evaluation statistics as Huang et al. (2014). Source to Interference Ratio (SIR), Source to Artifacts Ratio (SAR), and Source to Distortion Ratio (SDR) by BSS-EVAL 3.0 metrics are used by the python package `mir_eval`.

5. Results

5.1. Project Results

In our limited time of training and tuning parameters, we have only achieved a GNSDR score of 6.13 for the original model structure. Owing to not tuning our parameters to the best performance, we only achieve a compelling 6.08 GNSDR score.

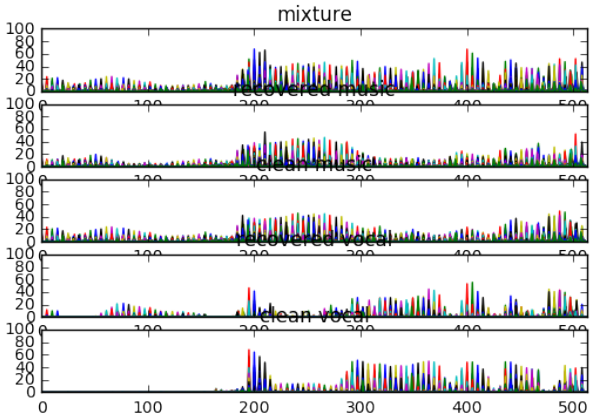

Figure 2. Time-Frequency Plot for original algorithm: (from top to bottom) mixture magnitude, clean vocal, recovered vocal, clean music and recovered music.

The respected magnitude plots are plotted in Fig. 2 and Fig. 3.

5.2. Comparison of Results

Though still not as compelling as the original results documented in the paper, which have their best GNSDR scores over 7 (DRNN-1, 2, 3), we only adopt three 150-unit layers and haven't fully tuned our parameters to the specific models. We believe the model structure with still has potential to beat the original algorithm with the two adaptation we would like to add.

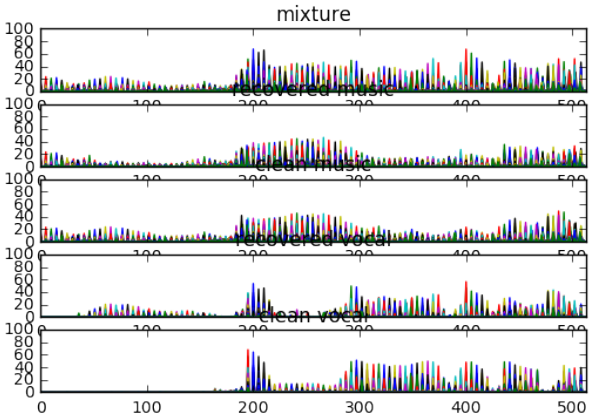

Figure 3. Time-Frequency Plot with IS-Optimizer: (from top to bottom) mixture magnitude, clean vocal, recovered vocal, clean music and recovered music.

5.3. Discussion of Insights Gained

Through this project, we have learned the following insights:

It is more difficult than it seems to build a neural network from scratch having every detailed dimensions calculated and sent, functions written and implemented. (without the usage of existing packages or toolkits built on Theano)

However, it is very worthwhile to follow the current trend of applying deep learning tools to powerfully or creatively solve existing problems, ranging from civil engineering, electrical engineering to art and literature.

We have learned this by studying how people use these new tools to do singing-voice separation specifically.

About our own project experience: even though we changed our topic in the middle, payed great efforts on the artistic style topic and came up with feasible strategies on that topic, without coding from scratch, we don't know the true commitment of mastering an existing subject. Debugging process can be extremely tedious when not fully understanding the model structures completely.

6. Conclusion and Future Work

In this project, we implement a functional neural network which separate singing-voice from the background music. Owing to the time constraint, we tune our parameters to achieve close GNSDR scores, which indicates the potential of the two possible extensions.

Given more time in the future, we would like to further tune our models and parameters and implement curriculum training.

References

[1] Code and Report. Bitbucket. https://bitbucket.org/e_4040_ta/e4040_project_hint

[2] Benesty, J., Sondhi, M. M., & Huang, Y. (Eds.). (2007). *Springer handbook of speech processing*. Springer Science & Business Media.

[3] Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009, June). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning* (pp. 41-48). ACM.

[4] Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016, March). Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 31-35). IEEE.

[5] Huang, P. S., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2014, October). Singing-Voice Separation from Monaural Recordings using Deep Recurrent Neural Networks. In *ISMIR* (pp. 477-482).

[6] Isik, Y., Roux, J. L., Chen, Z., Watanabe, S., & Hershey, J. R. (2016). Single-Channel Multi-Speaker Separation using Deep Clustering. *arXiv preprint arXiv:1607.02173*.

[7] Yang, P. K., Hsu, C. C., & Chien, J. T. (2014, October). Bayesian Singing-Voice Separation. In *ISMIR* (pp. 507-512).

Appendix

A.1 Individual student contributions - table

	cp2923	ks3403	sz2624
Last Name	Chang Pan	Kejia Shi	Shengyang Zhang
Fraction of (useful) total contribution	1/3	1/3	1/3
What I did 1	Main Coding and tuning	Report Writing / Part of Coding	Main Coding and tuning
What I did 2	Study for both papers	Study for both papers	Study for both papers